COMPARATIVE AND FUNCTIONAL GENOMIC ANALYSIS OF HUMAN AND CHIMPANZEE RETROTRANSPOSON SEQUENCES

A Dissertation
Presented to
The Academic Faculty

by

Nalini Polavarapu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biology

Georgia Institute of Technology
August 2007

COMPARATIVE AND FUNCTIONAL GENOMIC ANALYSIS OF HUMAN AND CHIMPANZEE RETROTRANSPOSON SEQUENCES

Approved by:

Dr. John McDonald, Advisor
School of Biology
Georgia Institute of Technology

Dr. James Thomas
Department of Human Genetics
Emory University

Dr. Jung Choi
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Date Approved: July 5, 2007

For my family

ACKNOWLEDGEMENTS

Throughout my graduate school years, I have been blessed to have the opportunity to meet people that have touched my life and from whom I have gained scientific and general knowledge which have allowed me to grow as an individual. First of all, I would like to thank my advisor, John F. McDonald, for his support and encouragement here at Georgia Tech and for future endeavors. I am grateful to the members of my committee for their advice, time and effort through the graduation process. Special thanks to King Jordan and Nathan Bowen for advice and input on my projects. Thanks to Marta Puig and Lilya Matyunina for teaching me lab work. Thanks to members of McDonald lab for support, laughter and making the lab a wonderful place to work. In particular, thanks to DeEtte Walker, Erin Dickerson, Laura Kapa and Gaurav Arora. I am indebted to Nina Schubert, Ahsan Huda and Lilya Matyunina for their awesome friendship, help and advice on all matters.

I would also like to thank my mother and father who have constantly encouraged me in my educational pursuits and helped me reach this amazing goal. This work would not have been possible without their love and support. My son Jai is a major driving force for my thesis. Missing him made me finish much faster than I would have otherwise. And thanks to the invisible driving force of my life whose presence made me travel paths I never imagined. Thanks to life. It's been a bumpy road but perfectly guided and directed.

There are many more people that are not named here that I have learned from and enjoyed being a small part of their lives. Thanks to you all!

TABLE OF CONTENTS

		Page
ACKNOWLEDGEMENTS		iv
LIST OF TABLES		viii
LIST OF FIGURES		x
LIST OF SYMBOLS AND ABBREVIATIONS		xi
SUMMARY		xiv
CHAPTER		
1	INTRODUCTION AND LITERATURE REVIEW	1
2	IDENTIFICATION, CHARACTERIZATION AND COMPARATIVE GENOMICS OF CHIMPANZEE ENDOGENOUS RETROVIRUSES	18
	Abstract	18
	Introduction	19
	Results and Discussion	20
	Conclusions	51
	Materials and Methods	56
	Acknowledgements	60
3	NEWLY IDENTIFIED FAMILIES OF HUMAN ENDOGENOUS RETROVIRUSES (HERVS)	61
	Introduction	61
	Results	61
	Acknowledgements	66

4 DIFFERENTIAL GENE EXPRESSION PATTERNS BETWEEN HUMANS AND CHIMPANZEES IS ASSOCIATED WITH RETROTRANSPOSON INDEL VARIATION 67

Abstract 67

Introduction 67

Results and Discussion 68

Methods 80

Acknowledgements 85

5 CONCLUSION 86

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER 2 90

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 4 94

APPENDIX C: PUBLICATIONS 126

REFERENCES 127

LIST OF TABLES

Page

Table 1.1: Transposable element content of sequenced genomes 5

Table 1.2: Examples of retrotransposon sequences functionally associated with host genes 6

Table 1.3: Examples of protein-coding host genes domesticated from retrotransposon sequences 9

Table 1.4: Some phenotypic traits of humans for comparison with those of great apes 11

Table 2.1: Previously characterized RT sequences from a variety of species used for comparison in phylogenies 22

Table 2.2: Representative sequences from each family of chimpanzee endogenous retroviruses 22

Table 2.3: Endogenous retrovirus INDEL sequences (>5000 bp) present in chimpanzees but absent in humans 44

Table 2.4: Endogenous retrovirus INDEL sequences (>5000 bp) present in humans but absent in chimpanzees 45

Table 2.5: Endogenous retrovirus INDELs (80 bp to 12.0 kb) in humans and chimpanzees 50

Table 3.1: Representative elements of human endogenous retrovirus families characterized in this study 63

Table 3.2: Characteristics of human endogenous retrovirus families identified in this study 64

Table 4.1: Categories of Human and Chimpanzee gap sequences 69

Table 4.2: Retrotransposon insertions and deletions in humans and chimpanzees 71

Table 4.3: Number of genes differentially expressed between humans and chimpanzees in different tissues 72

Table 4.4: Correlation (p-values) between INDEL variation and differences in human-chimpanzee gene expression patterns 73

Table 4.5: Indel variation associated with exons 74

Table 4.6: Correlation (p-values) between INDEL variation located in exon and intron of genes and human-chimpanzee differential gene expression 74

Table 4.7: Correlation (p-values) between INDEL variation located in upstream and downstream of genes and human-chimpanzee differential gene expression 75

Table 4.8: Human mRNA sequences associated with chimpanzee gaps 76

Table 4.9: Correlation (p-values) between retrotransposon indel variation and difference in human-chimpanzee gene expression 77

Table B.1: List of selected genes significantly differentially expressed in brain and associated with retrotransposon INDEL variation 94

Table B.2: List of selected genes significantly differentially expressed in testis and associated with retrotransposon INDEL variation 105

LIST OF FIGURES

Page

Figure 1: Classes are transposable elements (TEs) in human genome 4

Figure 2.1: Unrooted RT based neighbor joining tree of three classes of chimpanzee endogenous retroviruses 25

Figure 2.2: RT-based neighbor-joining tree for Class I chimpanzee endogenous retroviruses 29

Figure 2.3: RT-based neighbor-joining tree for Class II chimpanzee endogenous retroviruses 31

Figure 2.4: Insertion of a member of CERV 30 (HERVK10) family in chimpanzees 32

Figure 2.5: RT-based neighbor-joining tree for Class III chimpanzee endogenous retroviruses 33

Figure 2.6: Phylogenetic tree of CERV 1/PTERV1 LTRs 35

Figure 2.7: Unrooted neighbour joining phylogenetic tree built from solo LTRs and 5' and 3' LTRs of full length elements of CERV 1/PTERV1 family 36

Figure 2.8: Unrooted neighbour joining phylogenetic tree built from solo LTRs of CERV 1/PTERV1 family 37

Figure 2.9: Phylogenetic tree of CERV 2 LTRs 39

Figure 2.10: Distribution of CERV 2 elements among primates 41

Figure 2.11: Structure of endogenous retroviral INDEL sequences (> 5000 bp) in humans 46

Figure 2.12: Structure of endogenous retroviral INDEL sequences (> 5000 bp) in chimpanzees 47

Figure 3: Unrooted RT based neighbour joining tree of human endogenous retrovirus families 65

LIST OF SYMBOLS AND ABBREVIATIONS

ANOVA	Analysis of Variance
BaEV	Baboon Endogenous Virus
BLAST	Basic Local Alignment Search Tool
BLV	Bovine Leukemia Virus
CDS	Coding Sequence
CERV	Chimpanzee Endogenous Retrovirus
DNA	Deoxyribonucleic Acid
ENV	Envelope
ERV	Endogenous Retrovirus
FeFV	Feline Foamy virus
FeLV	Feline Leukemia Virus
Gag	Group Specific Antigen
GALV	Gibbon Ape Leukemia Virus
GB	Gigabytes
GH-G18	Golden hamster intracisternal A-particle H18
HERV	Human Endogenous Retrovirus
HFV	Human Foamy Virus
HG16	Human Genome Version 16
HG18	Human Genome Version 18
HIV	Human Immunodeficiency Virus
INDEL	Insertion and Deletion
IV	Indel Variation

KoRV	Koala type C endogenous retrovirus
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MAS	MicroArray Suite
MDEV	Mus dunni endogenous retrovirus
MEGA	Molecular Evolutionary Genetics Analysis
MER	Medium reiteration frequency
MMTV	Mouse Mammary Tumor Virus
MuLV	Moloney murine Leukemia Virus
MY	Million Years
MYA	Million Years Ago
NCBI	National Center for Biotechnology Information
NR	Non-redundant database
ORF	Open Reading Frame
PBS	Primer Binding Site
PCR	Polymerase Chain Reaction
PERV	Porcine Endogenous Retrovirus
PPT	Poly Purine Tract
PTERV	Pan troglodytes Endogenous Retrovirus
RefSeq	Reference Sequence database
RERV	Rabbit Endogenous Retrovirus
RIV	Retrotransposon Indel Variation
RNA	Ribonucleic Acid
RSV	Rous Sarcoma Virus
RT	Reverse Transcriptase

SINE	Short Interspersed Nuclear Element
SRV	Simian Retrovirus
SVA	SINE R, VNTRs and Alu elements
TBLAST	Translated Basic Local Alignment Search Tool
T-Coffee	Tree-based Consistency Objective Function for Alignment Evaluation
TE	Transposable element
tRNA	transfer RNA
TSD	Target Site Duplication
TSR	Target Site Repeat
UCSC	University of California Santa Cruz
VNTR	Variable Number of Tandem Repeats

SUMMARY

Transposable elements (TEs) are mobile DNA sequences that can move from one location to another in the genome. These elements encode regulatory features including transcriptional promotion and termination signals facilitating the production of new transcripts (or elements). The elements thus produced are inserted back into the genome. Due to their insertional capacity and encoded regulatory features, TEs have, in recent years, been recognized as significant contributors to regulatory variation both within and between species. In comparing the human and chimpanzee genomes it has been hypothesized that the genetic basis of the phenotypic differences that distinguish them may be the result of regulatory differences existing between the two species. Since TEs inserted in proximity to genes can significantly alter gene expression patterns, this research aims at exploring the influence of TE sequences and retrotransposons in particular in the evolution of gene regulation between humans and chimpanzees.

High-throughput genome processing, microarray analysis, programming, statistical analysis, information management and experimental evidence were used to achieve two major advances in the comparative genomics of human and chimpanzee genomes.

Research Advance 1: A first systematic search of one particular class of retrotransposons - endogenous retroviruses (ERVs) was carried out in the chimpanzee genome. Forty two families of ERVs were identified in the chimpanzee genome including the discovery of 9 previously unknown families in humans. The vast majority of these families were found

to have orthologs in the human genome except for two (CERV 1/PTERV1 and CERV 2) families. The two CERV families without orthologs in the human genome display a patchy distribution among primates. Nine families of chimpanzee ERVs have been transpositionally active since the human-chimpanzee divergence, while only two families have been active in the human lineage.

Research Advance 2: The genomic differences [INDEL variation (80-12,000 bp in length)] between humans and chimpanzees are laid out. The INDEL variation located in or near genes is categorized in detail and is correlated with differences in gene expression patterns in a variety of organs and tissues. Results indicate that the majority of the INDEL variation between the two species is associated with retrotransposon sequences and that this variation is significantly correlated with differences in gene expression most notably in brain and testes. These findings are consistent with the hypothesis that retrotransposon mediated regulatory variation may have been a significant factor in human/chimpanzee evolution.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION

Introduction to Transposable Elements (TEs)

Transposable elements (TEs) are mobile DNA sequences that move from one location to another in the genome. They were first postulated in the 1940s by Barbara McClintock in the maize genome (McClintock 1946; McClintock 1984) and experimentally verified in the 1960s in bacteria. Since then, they have been found to be associated with a variety of types of mutations (e.g. insertions, deletions and translocations) in several organisms including humans [e.g. (Kazazian 1998; Deininger and Batzer 1999)]. Transposable elements are grouped into two major classes based on their mode of transposition (Finnegan 1992). Class I elements are retrotransposons that move in the genome by a "copy and paste" mechanism *via* an RNA intermediate and the element encoded enzyme reverse transcriptase (RT). The host transcription machinery transcribes the elements producing RNA which is then reverse transcribed into DNA by proteins produced from the translation of part of the RNA. The double stranded DNA thus produced is inserted into the genome. Class II TEs are DNA transposons that move in the genome by a "cut and paste" mechanism. The TE encoded enzyme "transposase" excises the element and inserts it elsewhere in the genome.

Class I elements are sub-divided into two groups: long terminal repeat (LTR) retrotransposons/endogenous retroviruses and non LTR retrotranposons consisting of

long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LTR retrotransposons/endogenous retroviruses and LINEs are autonomous retrotransposons, encoding reverse transcriptase (RT), the enzyme required for their reverse transcription. SINEs are non-autonomous retrotransposons that do not encode RT and utilize RT from LINEs (Boeke and Stoye 1997).

Evolutionary impact of transposable elements on host genomes

In their initial discovery by Barbara McClintock, transposable elements were described as controlling elements causing chromosomal breakage in response to stress (McClintock 1948; McClintock 1984). Based on the evidence, it was theorized that transposable elements may provide potential adaptive advantages to a host organism in the form of genome restructuring, and that their presence may therefore be maintained over time (McClintock 1951; Shapiro 1977; McClintock 1984). However concrete the results were, since then there has been much speculation about their evolutionary origins and the possible roles these elements play for their host genomes. In line with the 'phenotypic paradigm' of neo-Darwinian theory which holds that genes ensure their survival and representation in subsequent generations by providing selective advantage for their host organism (Doolittle and Sapienza 1980), it was speculated that the presence of transposable elements were supposed to be due to some function they perform for their host genomes. This was challenged in 1980 by two seminal papers published simultaneously in Nature (Doolittle and Sapienza 1980; Orgel and Crick 1980) leading to the selfish DNA theory of transposable elements. According to this theory, the emergence and spread of transposable elements could be explained solely by their ability to replicate

themselves in the genome. Because they can out-replicate and could even spread and persist in natural populations in the face of a selective disadvantage for their host organisms (Hickey 1982), their evolutionary success is largely irrelevant to any selective advantage they provide to their host genomes. Therefore, it was concluded that transposable elements are merely genomic parasites and their selfish nature precludes them from playing an important role in genome evolution.

However, findings in molecular biology and genomics accumulated indicating that an entirely selfish view of transposons is shortsighted and that they have been co-opted many times and in a number of different ways to serve the interests of their hosts. This led to the reemergence of an adaptive role for transposons in genomes (McDonald 1993, 1995; Brosius 1999b; Kidwell and Lisch 2001). Transposons have been shown to be important contributors to yeast double-strand break repair (Garfinkel 1997), *Drosophila* telomere maintenance (Pardue et al. 1996), and have also been implicated in mammalian DNA repair (Morrish et al. 2002). More importantly, they are now widely accepted to have played an important role in the evolution of epigenetic mechanisms. Epigenetic gene silencing mechanisms are now believed to have evolved initially to repress the activity of transposable elements. Subsequently epigenetic mechanisms are believed to have been co-opted by the hosts as primary epigenetic regulatory machinery in eukaryotic genomes. Even more remarkably, the presence of these epigenetic mechanisms may have facilitated the regulation of increased gene numbers associated with two major macroevolutionary transitions (Bird 1995). In this way, the selective pressure exerted on genomes by transposable elements played an important role facilitating two major

evolutionary transitions that are marked by considerable increases in genome complexity (McDonald 1998; Bowen and Jordan 2002).

Biological significance of retrotransposons, a major class of transposable elements

Retrotransposons are the most abundant and wide-spread class of eukaryotic transposable elements. For example, > 35 % of the mouse genome (Waterston et al. 2002), > 55% of the maize genome (SanMiguel et al. 1996) and > 45% of the human genome (Lander et al. 2001) are comprised of retrotransposon sequences (Table 1.1). The abundance and characteristic features of different classes of retrotransposons in the human genome are given in Figure 1. The biological significance of retrotransposons ranges from their contribution to mutation [e.g., (Green 1988)] and disease [e.g., (Kazazian 1998; Deininger and Batzer 1999)] to their role in gene and genome evolution [e.g., (McDonald 1993; Britten 1996; Brosius 1999b)] . There are now hundreds of examples of retrotransposons or fragments of retrotransposons being functionally associated with genes [e.g., (Brosius 1999b); (Makalowski 2000; Medstrand et al. 2001)] (Table 1.2).

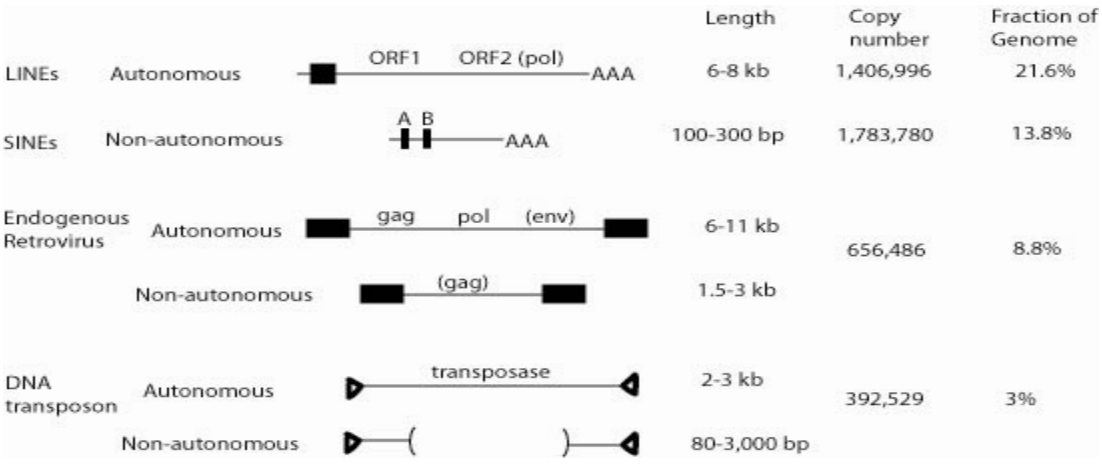

Figure 1: Classes are transposable elements (TEs) in human genome (Lander et al. 2001)

Table 1.1 Transposable element content of sequenced genomes

Scientific Name	Organism	Genome size (Mbp)	%TE	Source
Anopheles gambiae	Mosquito	278	>20%	(Holt et al. 2002)
Arabidopsis thaliana	Arabidopsis	130	~10%	(Le et al. 2000)
Bos taurus	Cow	3,247	>46%	(Larkin et al. 2003)
Caenorhabditis elegans	C. elegans	103	6%	(Consortium 1998)
Candida albicans	Human pathogen	16	~1.0%	(Goodwin and Poulter 2000)
Canis familiaris	Dog	2,384	34%	(Lindblad-Toh et al. 2005)
Ciona intestinalis	Sea squirt	153	12%	(Satoh et al. 2003)
Drosophila melanogaster	Fruitfly	132	15%	(Hoskins et al. 2002; Kaminker et al. 2002)
Fugu rubripes	Pufferfish (marine)	393	<15%	(Aparicio et al. 2002)
Gallus gallus	Chicken	1,050	<9%	(Consortium 2004)
Homo sapiens	Human	3,200	>45%	(Li et al. 2001)
Hordeum vulgare	Barley	4,800	>51%	(Vicient et al. 2001; Rostoks et al. 2002)
Lillium	Lilies	36,000	>95%	(Leeton and Smyth 1993)
Macaca mulatta	Rhesus monkey	3,093	~50%	(Han et al. 2007)
Monodelphis domestica	Opossum	3,501	>52%	(Mikkelsen et al. 2007)
Mus musculus	Mouse	3,000	38%	(Waterston et al. 2002)
Neurospora crassa	Bread mold	40	~8%	(Selker et al. 2003)
Oryza sativa	Rice	430	<25%	(Bennetzen 2000)
Pan troglodytes	Chimpanzee	2,928	>45%	(Mikkelsen et al. 2005)
Pinus taeda	Pine	~20-30,000	>50%	(Friesen et al. 2001)
Rattus norvegicus	Rat	2,507	~40%	(Gibbs et al. 2004)
Saccharomyces cerevisiae	Baker's yeast	12	3.1%	(Kim et al. 1998)
Schizosaccharomyces pombe	Fission yeast	12.2	1.1%	(Bowen et al. 2003)

Table 1.1 continued

Scientific Name	Organism	Genome size (Mbp)	%TE	Source
Tetraodon nigroviridis	Pufferfish (freshwater)	342	<10%	(Dasilva et al. 2002)
Triticum aestivum	Wheat	16,000	>90%	(Flavell 1986)
Zea mays	Corn	~2,300-3,000	>55%	(Kumar and Bennetzen 1999; Meyers et al. 2001)

Retrotransposons harbor polyadenylation signals, promoter and enhancer sequences which may significantly alter the expression patterns of the genes when integrated in their proximity. There are now several examples of the presence of retrotransposon sequences near the genes resulting in a change in gene expression [e.g.(Banville and Boie 1989; Di Cristofano et al. 1995; Hamdi et al. 2000; Medstrand et al. 2001; Nigumann et al. 2002; Kashkush et al. 2003; Gerlo et al. 2006)] or coding potential (Murnane and Morales 1995) without destroying existing gene functions (Table 1.2). Retrotransposon sequences are used as alternate promoters [e.g. (Yang et al. 1998; Medstrand et al. 2001)], enhancers [e.g. (Yang et al. 1998)], polyadenylation signals (Harendza and Johnson 1990) and alternatively spliced exons (Sorek et al. 2002) for several genes (Table 1.2). Interestingly, retrotransposon genes have evolved as new genes with functions beneficial to the host (Zdobnov et al. 2005; Volff 2006), a phenomenon termed as "molecular domestication" which is believed to have occurred repeatedly during the evolution of eukaryotic genomes (Miller et al. 1997; Miller et al. 1999) (Table 1.3).

Table 1.2 Examples of retrotransposon sequences functionally associated with host genes

Retrotransposon	Associated gene	Organism	Function	Source
LTR	Aromatase	Chicken	Promoter and 5' exon	(Matsumine et al. 1991)
LINE	Bcnt	Cow	Endonuclease	(Iwashita et al.

Table 1.2 continued

Retrotransposon	Associated gene	Organism	Function	Source
			domain	2003)
LTR	Chitinase 3	Drosophila	Potential enhancer	(McCollum et al. 2002)
LTR	White locus	Drosophila	Insulator	(Conte et al. 2002)
LTR	White locus	Drosophila	Enhancer	(Conte et al. 2002)
LTR	Cyp6g1	Drosophila	5' UTR, DDT resistance	(Daborn et al. 2002)
LTR	Cadherin-superfamily	Heliothis virescens (cotton pest)	Bt pesticide resistance	(Gahan et al. 2001)
Alu	Interferon-gamma	Human/primate	Gene regulation	(Ackerman et al. 2002)
Alu	Pax6 (transcript factor)	Human	Binding site of Pax6	(Zhou et al. 2002)
Alu	Cathepsin B	Human	Exon 2	(Berquin et al. 1997)
Alu	Human hematopoietic progenitor kinase (HPK1)	Human	Extension of coding sequence	(Hu et al. 1996)
Alu	Adenosine deaminase	Human	Extension of coding sequence	(Gerber et al. 1997)
LTR	Zinc finger gene	Human	Promoter	(Di Cristofano et al. 1995)
LTR	Apolipoprotein C-I	Human	Alternative promoter	(Medstrand et al. 2001)
LTR	8-oxo-dGTPase	Human	Alternative start	(Oda et al. 1997)
LTR	cDNA 7, cDNA y	Human	Poyadenylation signal	(Paulson et al. 1987)
LTR	AF-3	Human	Promoter	(Feuchter et al. 1992)
LTR	PLT	Human	Polyadenylation signal	(Goodchild et al. 1992)
LTR	cH-6	Human	Polyadenylation signal	(Mager 1989)
LTR	Calbindin D28K	Human	Promoter	(Liu and Abraham 1991)
LTR	Gin-1	Human/mammals	Gypsy-like integrase domain	(Llorens and Marin 2001)
LTR	MyEF-3	Human / mouse	Exon - domains	(Volff et al. 2001)

Table 1.2 continued

Retrotransposon	Associated gene	Organism	Function	Source
LTR	Endothelin B receptor	Human	Alternative tissue specific promoter (placental)	(Medstrand et al. 2001)
LTR	Amylase	Human	Tissue-specific promoter (parotid)	(Ting et al. 1992)
LTR	Erythroid beta-globin locus control region (beta-LCR)	Human	Alternative tissue specific promoter (embryonic and hematopoietic cells)	(Ling et al. 2002)
LTR	Leptin receptor	Human	Alternative splicing / termination	(Kapitonov and Jurka 1999)
LTR	HHLA3	Human	Polyadenylation signal	(Mager et al. 1999)
LTR	mid1	Human	Tissue specific potential enhancer	(Landry et al. 2002)
LTR	Pituitary hormone prolactin (PRL)	Human	Alternative promoter	(Gerlo et al. 2006)
LINE	Apolipoprotein (a)	Human	Enhancer	(Yang et al. 1998)
LINE	MET- proto oncogene	Human	Alternative promoter, exons	(Nigumann et al. 2002)
LINE	TACTILE (T-cell surface antigen)	Human	Alternative promoter, exons	(Nigumann et al. 2002)
LINE	SPT3	Human	Alternative promoter, exons	(Speek 2001)
LINE	Methyl-CpG binding protein	Human	Alternative stop	(Yu et al. 2001)
LINE	Thymidylate synthase	Mouse	Polyadenylation signal	(Harendza and Johnson 1990)
LTR	Sex-limited protein (slp)	Mouse	Promoter	(Stavenhagen and Robins 1988)
LTR	MIPP	Mouse	Promoter	(Chang-Yeh et al. 1991)
LTR	A1	Mouse	Polyadenylation signal	(Baumruker et al. 1988)
LTR	Oncomodulin	Rat	Promoter	(Banville and Boie 1989)

Table 1.2 continued

Retrotransposon	Associated gene	Organism	Function	Source
LINE	Insulin I gene	Rat	Transcriptional silencer	(Laimins et al. 1986)
LTR	Xa21 (disease resistance)	Rice	Regulation	(Richter and Ronald 2000)
LTR	Genes w/ Pol III promoters	Yeast	Possible Upregulation	(Bolton and Boeke 2003)

Table 1.3 Examples of protein-coding host genes domesticated from retrotransposon sequences [Information obtained from (Volff 2006)]

Gene (family)	Protein properties and function	Ancestral TE gene	Organism	Source
Iris	Defence against viruses?	Env	Drosophila	(Malik and Henikoff 2005)
Telomerase	Ribonucleoprotein, reverse transcriptase, telomere replication	RT	Eukaryotes	(Lingner et al. 1997; Nakamura et al. 1997)
Syncytin-1	Membrane glycoprotein, cell fusion, placenta formation	Env	Human	(Mi et al. 2000)
Syncytin-2	Membrane glycoprotein, cell fusion, placenta formation	Env	Human	(Blaise et al. 2003)
Peg10/Mart2	Cell proliferation, transcription factor?, placenta formation	Gag	Mammals	(Ono et al. 2006)
Ldoc1/Mart7	Inhibition of NF-kappaB activation, induction of apoptosis	Gag	Mammals	(Nagasaki et al. 2003)
Map-1/Ma1	Bax-associating protein, induction of apoptosis	Gag	Mammals	(Dalmau et al. 1999; Tan et al. 2001)
OtherMa genes	Neuronal autoantigens in paraneoplastic neurological diseases	Gag	Mammals	(Wills et al. 2006)
Iris-like	Defence against viruses?	Env	Mosquitoe	(Malik and Henikoff 2005)
Fv1	Murine leukemia virus restriction	Gag	Mouse	(Best et al. 1996)
Syncytin-A/B	Membrane glycoprotein, cell fusion, placenta formation	Env	Mouse	(Dupressoir et al. 2005)

In humans, at least 4% of protein-coding regions (Nekrutenko and Li 2001), 25% of promoter regions (Jordan et al. 2003) and 27% of untranslated regions of the genes contain identifiable retrotransposon and derived sequences (van de Lagemaat et al. 2003). In cases where they have been studied, these sequences have been found to affect the expression of genes through the contribution of transcriptional regulatory signals (van de Lagemaat et al. 2003; Thornburg et al. 2006) and are postulated to have played a significant role in the diversification and evolution of mammalian genes, particularly, in the evolution of human gene regulation. Indeed, it is now widely believed that an understanding of the origins and biological significance of retrotransposons and other TEs is prerequisite to a full understanding of the evolution of gene, genome structure and function (Li et al. 2001).

Comparison with the chimpanzee genome is ideal for the analysis of the contribution of retrotransposons to human gene evolution

Although humans and chimpanzees have accumulated significant differences in a number of morphological, behavioral, cognitive and phenotypic traits (Varki and Altheide 2005) (Table 1.4) since diverging from a common ancestor about six million years ago, their genomes are > 98.5% identical at protein coding loci (Mikkelsen et al. 2005). Since this modest degree of nucleotide divergence does not seem sufficient to explain the extensive phenotypic differences that exist between the two species (Varki and Altheide 2005) (Table 1.4), it has been hypothesized that the genetic basis of the differences lies at the level of gene regulation (King and Wilson 1975; Carroll 2003). Direct evidence in support of the regulatory hypothesis has recently been provided by a number of

comparative microarray studies showing that significant differences in gene expression patterns exist between humans and chimpanzees especially in organs (e.g., brain and testes) and functions (e.g., cognitive ability and fertility) directly related to some of the major phenotypic traits distinguishing the two species [e.g. (Enard et al. 2002; Caceres et al. 2003; Gu and Gu 2003; Khaitovich et al. 2005)]. The question remains, however, as to what is the genetic basis of the differences in gene regulation that separates humans from chimpanzees.

Table 1.4 Some phenotypic traits of humans for comparison with those of great apes[a] [adapted from (Varki and Altheide 2005)]

Life History	**Organ Physiology**	**Nutrition**	**Cognitive capacity**
Secondary Altriciality	Aldosterone Response to Posture	Frugivory	Declarative Memory
Helplessness of the Newborn	Salt-Wasting Kidneys	Carnivory	Imitative Learning
Prolonged Helplessness of Young	Ability For Sustained Running	Aquatic Foods	Teaching
Extended Care of Young	Voluntary Control of Breathing	Underground Foods	Symbolic Representation
Childhood	Ability to Dive Underwater	Cooking	Awareness of Death
Adolescence	Diving Reflex		Awareness of the Past
Age at First Reproduction	Ability to Float/Swim	**Neuroanatomy**	Awareness of the Future
Longevity	Emotion Lacrimation	Relative Brain Size	Theory of Mind
	Salt Content of Tears	Direct Cortical Projections	Theory of Other Minds
Reproductive Biology	Olfactory Sense	Relative Volume of Frontal Cortex	Empathy
Concealed Ovulation		Relative Volume of Corpus Callosum	Numeracy
Virgin Breast Development	**Cell Biology**	Relative Volume of Cerebellum	
Female Pituitary Menopause	No Differences Are Known?	% of Brain Growth Complete at Birth	**Communication**

Table 1.4 continued

Placentophagy		Rate of Postnatal Brain Growth	"Parentese" Sounds
Female Labia Majora	**Biochemistry**		Infant "Protoconversations"
Vaginal Hymen	Placental Alkaline Phosphatase	**Neurobiology**	Gestural Communication
Baculum (Penis Bone) Sperm Count	N-Glycolylneuraminic Acid Expression	Population Distribution of Handedness	Symbolic Communication
	Alpha 2-6-Linked Sialic Acid Expression		Semantics
Copulatory Plug		Postnatal Dendritic	Grammar and Syntax
		Growth	
		Postnatal Synapse Formation	
	Endocrinology		Recursion
Embryology	Thyroid Hormone Metabolism	Cortical Synapse Density	Writing
		Cortical Neuron Density	
Early Fetal Wastage/Aneuploidy	**Pharmacology**	Dendrites Per Neuron	**Social Organization**
Hydatiform Molar Pregnancy	Methylation of Inorganic Arsenic	Synapses Per Neuron	
Umbilical Cord Length		Adult Neurogenesis	Institutions
		Cingulate Cortical Spindle Neurons	
	Anatomic pathology		Social Conventions
	Cortical Neurofibrillary Tangles	Finger Tip Sensory Nerve Endings	
			Governments
Pregnancy/Parturition	**Clinical Pathology**	**Neurochemistry**	Enforcement Through Sanctions
Cephalo-pelvic Disproportion	Erythrocyte Sedimentation Rate	Brain Aromatisation of Testosterone	
Duration of Labor			
Maternal Mortality in Childbirth			
Pain During			

Table 1.4 continued

Childbirth			
Need for Assistance with	Serum Alkaline Phosphatase Level	Tyrosine Hydroxylase Heterogeneity	**Culture**
Childbirth	RBC and Serum Folate		Composition of Art
Neonatal Cephalhematoma	Serum Vitamin B12/B12 Binding	**Mental Disease**	Composition of Music
	Total Leukocyte Count	Schizophrenia	Composition of Rhythms
Postnatal Development	Absolute Neutrophil Count	Bipolar Psychosis	Death Rituals
Late Closure of Cranial Sutures	Absolute Lymphocyte Count	Autism	Clothing (Covering of
Duration of Infant Arousal		Suicide	Body Parts)
Inconsolable Infant Crying	**Dental Biology/Disease**		Rites of Passage
Infant-Caregiver Attunement	Canine Tooth Diastema	**Behavior**	Genocide
Maternal-Infant Eye-To-Eye Gaze	Canine Tooth Dysmorphism	Control of Facial Expressions	Competitive Sports
	Tooth Enamel Thickness	Planning Ahead	Practicing of Skills
Anatomy	Retromolar Gap	Intentional Deception	Physical Modifications of the Body
Sagittal Crest of Skull	Third Molar Impaction	Deliberately Delaying Gratification	
Brow Ridge	Dental Eruption Sequence/Timing	Long-Range Transport of Materials	Inheritance of Resources and Status
Protuberantia Menti (Chin)		Secondary Tool-Making	
Length of Sphenoid Sinus	**Medical/Surgical Diseases**	Mechanical Multi-Tasking	Rythmic Dance
Choroid Plexus Biondi Bodies	HIV Progression to AIDS	Physical Abuse of the Young	Sculpture
Inner Ear Canal Orientation	*P. falciparum malaria*	Torture	Belief in Supernatural/ Religion
Apical Phalangeal Tufts	Viral Hepatitis B/C Complications	Organized Warfare	
Age of Pelvic Bone	Influenza A	Adult Play	Body Adornment

Table 1.4 continued

Fusion	Infection Severity		
Bone Cortex Thickness	Incidence of Carcinomas	Symbolic Play	Childbirth Customs
Laryngeal Position	Hemorrhoids	Abuse of Other Animals	Sexual Intercourse in Private
Pharyngeal Air Sacs	Varicose Veins	Inter-Group Coalition Formation	Gift-Giving
Ear Lobes	Pelvic Phleboliths	Use of Containers	Hospitality
Sexual Body Size Dimorphism	Foamy Virus (Spumavirus) Infections	Care of Infirm and Elderly	Intertwining (e.g., weaving)
Lacrimal Gland Structure	Sexually Transmitted Diseases	Grandparenting	
Visible Whites of the Eyes		Home Base	Meal Times
Small/Large Intestine Length Ratio	**Immunology**	Control of Fire	Poetry
Meningeal Artery Source	Sialoadhesin on Macrophages	Food Preparation	Property
		Organized Gathering of Food	Construction of Shelters
Biomechanics	**Skin Biology and Disease**	Domestication of Animals	Taboos
Bipedal Gait	Eyebrows	Domestication of Plants	Taxonomy of Species
Adductive Thumb	Eccrine Sweat Glands	Altruistic Punishment	Trade
Skeletal Muscle Strength	Acne Vulgaris	Peace-Making	Measurement of Time
Hand-Eye Coordination	Subcutaneous Fat	Somnambulism	Weapons
Fine Motor Coordination	Body Lice	Mind-Altering Drug Use	Toys

[a]A major limitation in translating genomic comparative information into an understanding of "humanness" is that we know relatively little about the basic phenotypic features of the great apes, relative to humans. This table lists topic areas in which there are real or claimed "differences" between humans and the great apes (as a group). A given "difference" listed here could be a suggested gain or loss in humans, with respect to the great apes.

One recently offered hypothesis is that the substantial INDEL (insertion/deletion) variation that exists between humans and chimpanzees may contribute significantly to the

regulatory differences between the species [e.g. (Britten 2002; Frazer et al. 2003; Newman et al. 2005)]. The INDEL variation can occur due to the transposition events of retrotransposon sequences leading to their accumulation in the genomes together with the elimination of these sequences from the genomes. It has been estimated that the human genome has expanded 20% over the last 50 MY, almost entirely due to retrotransposon insertions (Liu et al. 2003). Since retrotransposon sequences located in or near genes are known to have the ability to significantly alter patterns of gene expression [e.g. (Hamdi et al. 2000; Kashkush et al. 2003)], these elements have been recognized as significant contributors to regulatory variation both within and between species [e.g., (Britten 1996; Brosius 1999a)]. Therefore, if one wants to fully comprehend the genomic differences between humans and chimpanzees, the differences among their retrotransposons must be laid out in full. The sequencing of the chimpanzee genome (Mikkelsen et al. 2005) and the availability of expression data from five different tissues (*viz* brain, heart, liver, kidney, testis) in humans and chimpanzees (Khaitovich et al. 2005) has provided an unprecedented opportunity to not only compare the full complement of retrotransposons in two closely related primate species but to gain insight into the role these elements may have played in human evolution, particularly, in the evolution of gene regulation between humans and chimpanzees.

Analysis of retrotransposon contribution to human-chimpanzee regulatory evolution

Previously, studies concerning the influence of retrotransposons in gene and genome evolution have focused most intensively on humans (Nekrutenko and Li 2001; Jordan et

al. 2003; van de Lagemaat et al. 2003). This dissertation extends the analysis to chimpanzee genome to analyze the regulatory influence of retrotransposons in two closely related primate species. An extensive computational approach was used to search the chimpanzee genome for endogenous retroviruses, a class of retrotransposons, followed by a comparative genomic approach to identify gene-associated retrotransposon INDEL variation (RIV) between the two species. The RIV was correlated with differences in gene expression patterns between humans and chimpanzees.

Chapter 2 presents the results of the first systematic search for endogenous retroviruses in the chimpanzee genome. Chimpanzee genome contains at least 42 separate families of endogenous retroviruses, nine of which were not previously identified. All but two (CERV 1/PTERV1 and CERV 2) of the 42 families of chimpanzee endogenous retroviruses were found to have orthologues i.e. elements in corresponding genomic positions, in the human genome. Nevertheless, nine families of chimpanzee ERVs have been transpositionally active since the human-chimpanzee divergence, while only two families have been active along the human lineage. The two CERV families without orthologues in the human genome display a patchy distribution among primates. A survey of endogenous retroviral positional variation between chimpanzees and humans determined that approximately 7% of all human-chimpanzee INDEL variation is associated with endogenous retroviral sequences.

Chapter 3 describes and characterizes in detail the nine previously unidentified families of Human Endogenous Retroviruses (HERVs). All are low abundance families being

comprised of only 1-7 full-length members with low homology to previously identified HERVs. Each of the newly identified families is represented by significantly more solo LTRs and fragmented sequences than full-length elements. The estimated age of these families range from 18.0 to 49.5 MY indicating that members of these families have not been transpositionally active in the primate lineage since well before chimpanzees and humans diverged from a common ancestor (6 MYA).

Chapter 4 is a detailed characterization of INDEL variation existing between human and chimpanzee genomes with particular emphasis on the variation associated with human and chimpanzee genes. The gene-associated INDEL variation is correlated with differences in the gene expression patterns in a variety of organs and tissues. The results indicate that the majority (~60%) of INDEL variation between humans and chimpanzees is associated with retrotransposon sequences and that this variation is significantly correlated with differences in gene expression patterns most notably in brain and testes. These results indicate that retrotransposon mediated regulatory variation may have been a significant factor in human/chimpanzee evolution.

CHAPTER 2

IDENTIFICATION, CHARACTERIZATION AND COMPARATIVE GENOMICS OF CHIMPANZEE ENDOGENOUS RETROVIRUSES

ABSTRACT

Background

Retrotransposons, the most abundant and wide-spread class of eukaryotic transposable elements, are believed to play a significant role in mutation and disease and to have contributed significantly to the evolution of genome structure and function. The recent sequencing of the chimpanzee genome provided an unprecedented opportunity to study the functional significance of these elements in two closely related primate species and to better evaluate their role in primate evolution.

Results

We report here that the chimpanzee genome contains at least 42 separate families of endogenous retroviruses; 9 of which were not previously identified. All but two (CERV 1/PTERV1 and CERV 2) of the 42 families of chimpanzee endogenous retroviruses were found to have orthologues in humans. Molecular analysis (PCR and Southern Hybridization) of CERV 2 elements demonstrates that this family is present in chimpanzee, bonobo, gorilla and old world monkeys but absent in human, orangutan and new world monkeys. A survey of endogenous retroviral positional variation between chimpanzees and humans determined that ~7% of all chimpanzee-human INDEL variation is associated with endogenous retroviral sequences.

Conclusion

Nine families of chimpanzee endogenous retroviruses have been transpositionally active since chimpanzees and humans diverged from a common ancestor. Seven of these transpositionally active families have orthologues in humans, one of which has also been transpositionally active in humans since the human – chimpanzee divergence ~6 MYA. Comparative analyses of orthologous regions of the human and chimpanzee genomes has revealed that a significant portion of INDEL variation between chimpanzees and humans is attributable to endogenous retroviruses and may be of evolutionary significance.

INTRODUCTION

Retrotransposons are the most abundant and wide-spread class of eukaryotic transposable elements. For example, > 35 % of the mouse genome (Waterston et al. 2002), > 55% of the maize genome (SanMiguel et al. 1996) and > 45% of the human genome (Lander et al. 2001) are comprised of retrotransposon sequences. This group of transposable elements is made up of short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs) and long-terminal-repeat (LTR) retrotransposons/endogenous retroviruses, all of which replicate via reverse transcription of an RNA intermediate (Boeke and Stoye 1997). The biological significance of retrotransposons ranges from their contribution to mutation [e.g., (Green 1988)] and disease [e.g., (Kazazian 1998; Deininger and Batzer 1999)] to their role in gene and genome evolution [e.g., (McDonald 1993; Britten 1996; Brosius 1999b)] .

The recent sequencing of chimpanzee genome has provided an unprecedented opportunity to not only compare the full complement of retrotransposons in two closely related primate species but to gain insight into the role these elements may have played in human evolution. We have combined the use of an LTR retrotransposon search algorithm, LTR_STRUC (McCarthy and McDonald 2003), with a systematic series of iterative TBLASTN searches to identify the endogenous retroviruses present in the Ensembl chimpanzee database (http://www.ensembl.org/Pan_troglodytes/). Since LTR_STRUC searches for LTR retrotransposons/endogenous retroviruses based on structure rather than homology, elements are often identified that go undetected in traditional BLAST searches [e.g., (McCarthy and McDonald 2003)].

LTR_STRUC is designed specifically to find full-length LTR retrotransposons/endogenous retroviruses, i.e., ones having two LTRs and a pair of target site duplications (TSDs) (McCarthy and McDonald 2003). Thus, we complemented our search by using reverse transcriptase (RT) sequences from LTR_STRUC-identified elements as query sequences in an iterative series of TBLASTN searches. This allowed us to identify structurally aberrant elements not directly detected by LTR_STRUC. Finally, a series TBLASTN searches were carried out using, as query sequences, previously reported human RT sequences for which orthologues were not identified by our previous two searches.

RESULTS AND DISCUSSION

The chimpanzee genome contains at least 42 families of endogenous retroviruses.

Using the procedure described above, we identified a total of 425 full-length chimpanzee endogenous retroviruses. This is certainly an under estimate of the number of endogenous retroviruses in the chimpanzee genome because we consciously excluded any sequences that could not be unambiguously identified as an endogenous retrovirus. The majority of these endogenous retroviruses (395/425 or 93%) were identified directly by LTR_STRUC or by homology to LTR_STRUC-identified elements.

ClustalX (Thompson et al. 1997) was used to build a multiple alignment of the RT domain of these 425 elements together with the RT domains of 16 previously described LTR retrotransposons/retroviruses representative of the three major classes of retroviral elements (Table 2.1). Phylogenetic analysis of the RT regions of the 425 full-length elements revealed the presence of at least 42 independent lineages of endogenous retroviruses in the chimpanzee genome that we here define as families (Figure 2.1). Non-autonomous endogenous retroviruses are elements that lack an RT ORF and are required to utilize RT activity from autonomous, full-length endogenous retrovirus in order to replicate. Many of the chimpanzee endogenous retrovirus families contain truncated, non-autonomous, as well as, full-length elements.

Forty of the 42 families of chimpanzee endogenous retroviruses identified in this study were found to have orthologues in the human genome including 9 that were identified in this study for the first time (Polavarapu et al. 2006a) (See Appendix A). Two previously identified chimpanzee endogenous retrovirus families do not have human orthologues (Table 2.2).

Table 2.1: Previously characterized RT sequences from a variety of species used for comparison in phylogenies (see figure 2.1, 2.2, 2.3, 2.5)

Name	Name of retrovirus	Accession number
RERV	Rabbit endogenous retrovirus	AF480925
GH-G18	Golden hamster intracisternal A-particle H18	GNHYIH
SRV-1	Simian SRV-1 type D retrovirus	M11841
MMTV	Mouse mammary tumor virus	NC_001503
RSV	Rous sarcoma virus	AF052428
HFV	Human foamy virus	Y07725
FeFV	Feline foamy virus	AJ223851
HIV-1	Human immunodeficiency virus 1	K03454
BLV	Bovine leukemia virus	K02120
BaEV	Baboon endogenous virus	X05470
FELV	Feline leukemia virus	M18247
MuLV	Moloney murine leukemia virus	AF033811
PERV	Porcine endogenous retrovirus	AF038601
MDEV	Mus dunni endogenous virus	AF053745
GALV	Gibbon ape leukemia virus	M26927
KoRV	Koala type C endogenous virus	AF151794

Table 2.2: Representative sequences from each family of chimpanzee endogenous retroviruses

ND: Not Determined
*Families submitted to Repbase

Table 2.2 continued

Family Name Chimpanzee family (orthologous human family)	tRNA primer	Location on Chromosome [chromosome no: position]	Target site repeats	Element length (bp)
CERV 1/PTERV1	Pro	8:62466629..62474817	GTAT/GTAT	8189
CERV 2	Pro	1:53871490..53880190	GTGA/GTGA	8338
CERV 3 (HERVS71)	Thr	7:45002408..45013133	AGGC/AGGC	10726
CERV 4 (HERV3)	Arg	6:65506183..65515842	TATA/TATA	9660
CERV 5 (HERV15)	Thr	20:22151622..22161290	TTTT/TTTT	9669
CERV 6 (HERV 1*)	Thr	8:43017710..43027603	CCAC/CCAC	9697
CERV 7 (Harlequin)	Glu	14:49265903..49274634	ATAAAT/ATAAAT	8745
CERV 8 (HERVE)	Glu	4:29336385..29342031	AACA/AACA	5647
CERV 9 (HERV 2*)	His	1:74420643..74424449	CTTTT/CTTTT	3807
CERV 10 (HERVH48)	Phe	1:162967211..16293841	ATTCT/ATTCT	6640
CERV 11 (HERV-H)	His	13:88231927..88241255	TGTTA/TGTTA	9329
CERV 12 (HERVFH19)	ND	1:9084130..9093376	ND	9236
CERV 13 (PRIMA4)	Arg	X:81351532..81361722	CCTC/CCTC	10191
CERV 14 (HERV 5*)	Leu, Arg	3:58020412..58027601	CACT/CACT	7198
CERV 15 (HERV-P)	Pro, Val	6:89330483..89339138	ATACC/ATACT	8500
CERV 16 (HERV-17)	Gln, Arg	4:55955342..55963662	CCTT/CCTT	8329
CERV 17 (HERV30)	Leu, Arg	5:121436599..121446300	AAAG/AAAG	9702
CERV 18 (HERV 9)	Arg, Lys, Pro	5:39234784..39242752	GGAG/GGAG	7966
CERV 19 (PABL_B)	Arg, Leu	7:75174020..75182429	AGAG/AGAG	8410
CERV 20 (HERVP71A)	Pro	X:121795847..121803454	TTTTC/TTTTC	7608
CERV 21 (HERV 4*)	Thr, Pro	2:14653540..14661848	ATGA/ATGA	8321
CERV 22 (HERV 6*)	Thr	16:84359558..84368114	ND	8557
CERV 23 (HERV 7*)	Pro	17:39042670..39052505	AGAC/AGAC	9836
CERV 24 (HERV 10*)	Pro	17:27621756..27630879	TAAT/TAAT	9132
CERV 25 (HERV 11*)	Pro	1:32316520..32325874	GCAAA/GCAAA	9480
CERV 26 (HERV 12*)	ND	2:171938873..171948150	AATT/ACTT	9279
CERV 27 (HERVI)	Lys	5:122623439..122630725	CAGT/CAGT	7287
CERV 28 (HERVIP10F)	Ala	23:41095392..41106316	TACT/TACT	10925
CERV 29 (HERVG25)	ND	6:151527383..151531731	ND	4349
CERV 30 (HERVK10)	Lys	10:4757815..4766975	ATTAT/ATTAT	9161
CERV 31 (HERVK14)	Lys	9:74609341..74617943	CAATG/CAATG	8603
CERV 32 (HERVK14C)	Lys	12:84993042..85001650	ND	8609
CERV 33 (HERVK(C4))	Lys	1:134530572..134538281	ATTAAG/ATTAAG	7710
CERV 34 (HERVK9)	Lys	X:58520547..58526579	GCCTAG/GCCTAG	6033
CERV 35 (HERVK13)	ND	19:41852728..41865530	ND	12803
CERV 36 (HERVK11D)	Lys	5:123901088..123908580	ATAAAT/ATAAAT	7493
CERV 37 (HERVK11)	Lys	2:81402412..81411338	ATAAAA/ATAAAA	8927
CERV 38 (HERVK3)	Lys	5:26871940..26880204	GGTAAA/GGTGAA	8265
CERV 39 (HERVK22)	Met	5:103484321..103492150	GTTCTT/GTTCTT	7830
CERV 40 (HERV S)	Ser	1:147219818..147226527	CCATC/CCATC	6710
CERV 41 (HERV16)	ND	X:103812023..103815002	ND	2980
CERV 42 (HERVL)	Leu	3:83740925..83746481	ATAAT/ATAAT	5547

Table 2.2 continued

Family Name Chimpanzee family (orthologous human family)	Location on Chromosome [chromosome no: position]	5' and 3' LTR % identity	Length of 5'/3' LTRs (bp)	Dinucleotides
CERV 1/PTERV1	8:62466629..62474817	99.7	409/409	TG/CA
CERV 2	1:53871490..53880190	98.8	497/486	TG/CA
CERV 3 (HERVS71)	7:45002408..45013133	90	528/524	TG/CA
CERV 4 (HERV3)	6:65506183..65515842	91.7	643/592	TG/CA
CERV 5 (HERV15)	20:22151622..22161290	90	495/500	TG/CC
CERV 6 (HERV 1*)	8:43017710..43027603	92	511/512	TG/CA
CERV 7 (Harlequin)	14:49265903..49274634	90	477/470	TG/CA
CERV 8 (HERVE)	4:29336385..29342031	92.3	354/355	TG/CA
CERV 9 (HERV 2*)	1:74420643..74424449	74.5	318/315	TG/CA
CERV 10 (HERVH48)	1:162967211..16293841	86	404/401	TG/CA
CERV 11 (HERV-H)	13:88231927..88241255	91	367/367	TG/CA
CERV 12 (HERVFH19)	1:9084130..9093376	88	412/421	ND
CERV 13 (PRIMA4)	X:81351532..81361722	86	627/617	TG/CA
CERV 14 (HERV 5*)	3:58020412..58027601	83	420/430	TG/CA
CERV 15 (HERV-P)	6:89330483..89339138	87	634/625	TG/CA
CERV 16 (HERV-17)	4:55955342..55963662	89	775/760	TG/CA
CERV 17 (HERV30)	5:121436599..121446300	89	698/700	TG/CA
CERV 18 (HERV 9)	5:39234784..39242752	87	423/449	TG/CA
CERV 19 (PABL_B)	7:75174020..75182429	83	671/667	TG/CA
CERV 20 (HERVP71A)	X:121795847..121803454	85	464/458	TG/CA
CERV 21 (HERV 4*)	2:14653540..14661848	91	361/363	TG/CA
CERV 22 (HERV 6*)	16:84359558..84368114	86	434/435	AG/CT
CERV 23 (HERV 7*)	17:39042670..39052505	88	582/575	TG/CA
CERV 24 (HERV 10*)	17:27621756..27630879	87	429/431	TG/CA
CERV 25 (HERV 11*)	1:32316520..32325874	84	613/621	TG/CA
CERV 26 (HERV 12*)	2:171938873..171948150	93	509/506	TG/CA
CERV 27 (HERVI)	5:122623439..122630725	89.9	497/506	CG/CA
CERV 28 (HERVIP10F)	23:41095392..41106316	95.6	494/496	TG/CA
CERV 29 (HERVG25)	6:151527383..151531731	84	220/226	ND
CERV 30 (HERVK10)	10:4757815..4766975	99.4	961/958	TG/CA
CERV 31 (HERVK14)	9:74609341..74617943	92	623/618	TG/CA
CERV 32 (HERVK14C)	12:84993042..85001650	92	583/583	TG/CA
CERV 33 (HERVK(C4))	1:134530572..134538281	94	541/547	TG/CA
CERV 34 (HERVK9)	X:58520547..58526579	93.4	510/508	TG/CA
CERV 35 (HERVK13)	19:41852728..41865530	83	812/818	ND
CERV 36 (HERVK11D)	5:123901088..123908580	91	865/874	TG/TA
CERV 37 (HERVK11)	2:81402412..81411338	95.7	1079/1079	TG/CA
CERV 38 (HERVK3)	5:26871940..26880204	95.2	428/431	TG/CA
CERV 39 (HERVK22)	5:103484321..103492150	85	477/497	TG/CA
CERV 40 (HERV S)	1:147219818..147226527	85	325/329	TG/CA
CERV 41 (HERV16)	X:103812023..103815002	ND	ND	ND
CERV 42 (HERVL)	3:83740925..83746481	82	445/458	TG/CA

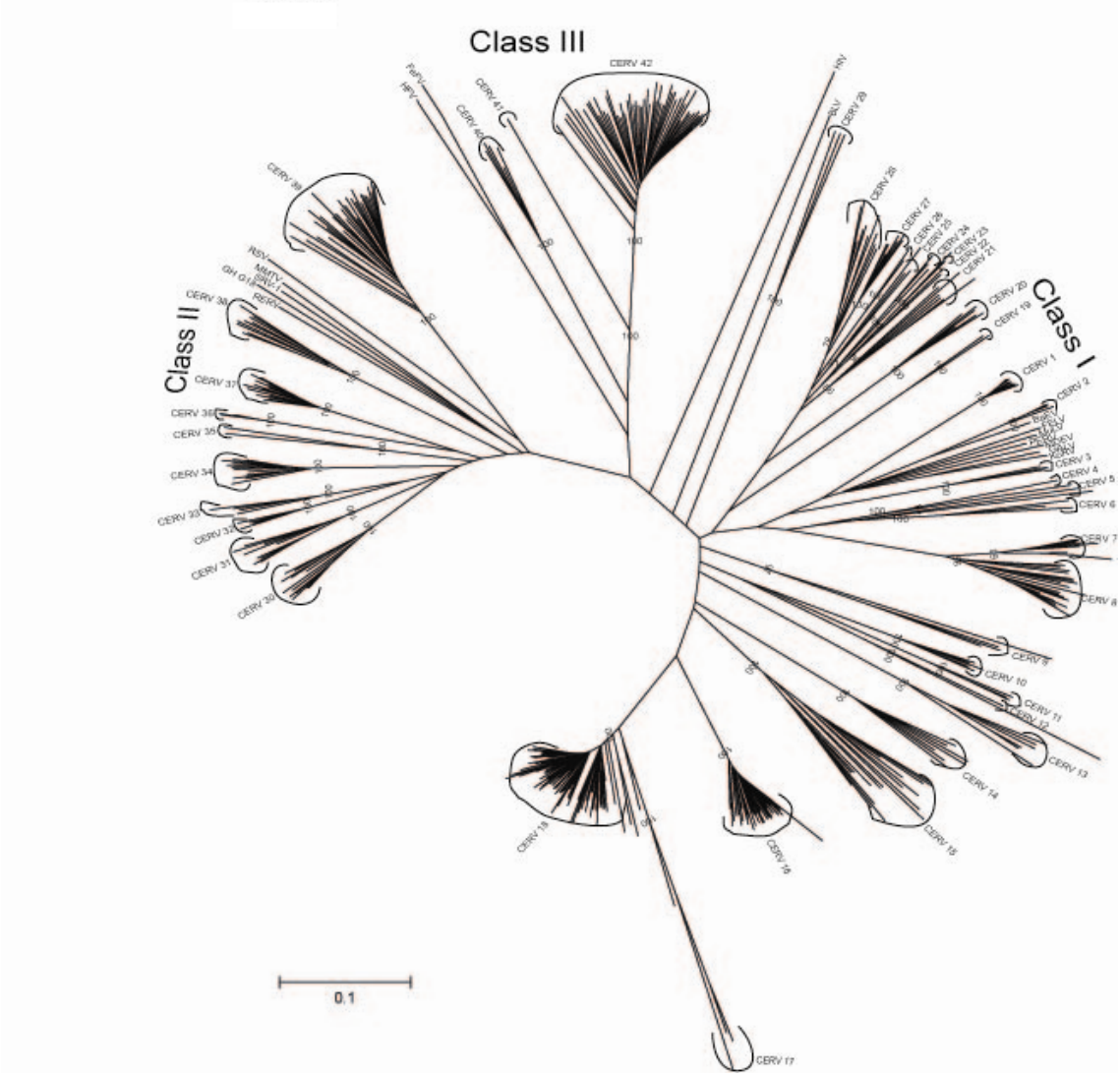

Figure 2.1 – Unrooted RT based neighbour joining tree of three classes of chimpanzee endogenous retroviruses
Class I: CERV1 to CERV29; Class II: CERV30 to CERV 39; Class III: CERV 40 to CERV 42. Bootstrap values are shown for each of the families. RT sequences from species other than chimpanzee, listed in table 2.1, are included for comparison

Consistent with the consensus nomenclature used for Human Endogenous Retroviruses (HERV) (Boeke and Stoye 1997), we here refer to the chimpanzee endogenous retroviral families by the acronym CERV (Chimpanzee Endogenous Retrovirus). Distinct families are indicated by number (e.g., CERV 1, ….CERV 42). In the single instance where the

CERV acronym refers to a previously named element/family, we include the pre-existing nomenclature as well (CERV 1/ PTERV1). In those cases where a CERV family has an orthologue in humans, the name of the orthologous HERV family is given in parentheses [e.g., CERV 3(HERVS71)].

Endogenous retroviral families of the chimpanzee genome

LTR retrotransposons and retroviruses are grouped into three major classes (Smit 1999). Class I contains elements related to the gammaretroviruses (e.g., MuLV : Moloney murine leukemia virus [Accesion No: AF033811], GALV : Gibbon ape leukemia virus [Accesion No: M26927] and FeLV : Feline leukemia virus [Accesion No: M18247]), Class II elements are related to betaretroviruses (e.g., MMTV : Mouse mammary tumor virus [Accesion No : NC_001503], RERV : Rabbit endogenous retrovirus [Accesion No : AF480925]) and Class III elements are distantly related to spumaviruses (e.g., HFV : Human Foamy virus [Accesion No : Y07725], FeFV : Feline Foamy virus [Accesion No : AJ223851]). Of the 42 chimpanzee families identified in our study, 29 belong to class I, 10 to class II and 3 to class III (Figure 2.1).

While there is a precedence for classifying human endogenous retroviruses into families based on their tRNA primer-binding sites [e.g., HERV K (lysine tRNA binding site)] (Boeke and Stoye 1997), we find that such groupings do not accurately reflect the phylogenetic groupings of CERVs. For example, some members of the CERV 21 family have a proline tRNA binding site whereas other members of this same family utilize threonine tRNA as a primer. Conversely, phylogenetically divergent CERV families may share the same tRNA binding site (e.g., members of the CERV 27 (HERV I) and CERV 30 (HERVK10) have lysine tRNA

binding sites) (Table 2.2). Thus, primer binding sites appear to be an evolutionarily labile feature and thus not a reliable indicator of phylogenetic relationships among chimpanzee endogenous retroviruses. A similar conclusion has been drawn for LTR retrotransposons in C. elegans (Ganko et al. 2001).

Full-length CERVs are typically between 7,000 and 10,000 bp in length. Consistent with what has been reported for LTR retrotransposons/endogenous retroviruses in other species (Bowen and McDonald 1999; McCarthy et al. 2002; McCarthy and McDonald 2004), CERV target site duplications (TSDs) range in size from four to six bp in length. With the exception of a few mutated copies, CERVs have the same canonical dinucleotides terminating the LTRs as have been reported for LTR retrotransposons/endogenous retroviruses in other species (TG/CA) (Bowen and McDonald 1999; McCarthy et al. 2002; McCarthy and McDonald 2004). CERV LTRs are typically 400-600 bp in length, although some LTRs are variant in size due to INDELs. For example, the LTRs of a member of the CERV 4 (HERV 3) family are 1591 bp in length due to the insertion of an *Alu* element at some point in the evolutionary history of this lineage. The following is a more detailed characterization of the 3 classes of CERVs.

Class I (families 1 - 29): The CERV families 1 through 29 group with the Class I retroviruses (Figure 2.1, Figure 2.2). The average size of full-length Class I CERVs is 8443 bp. These elements range in size from 2,268 – 13,135 bp in length. Much of this variation is due to INDELs associated with non-functional elements. The average size of LTRs associated with full-length Class I CERV elements is 544 bp (range 195 – 1591 bp). Class I CERV elements display considerable variation in their tRNA binding sites-

even within families (Table 2.2). The most frequently used tRNA primer for Class I CERV families (28%) is proline tRNA.

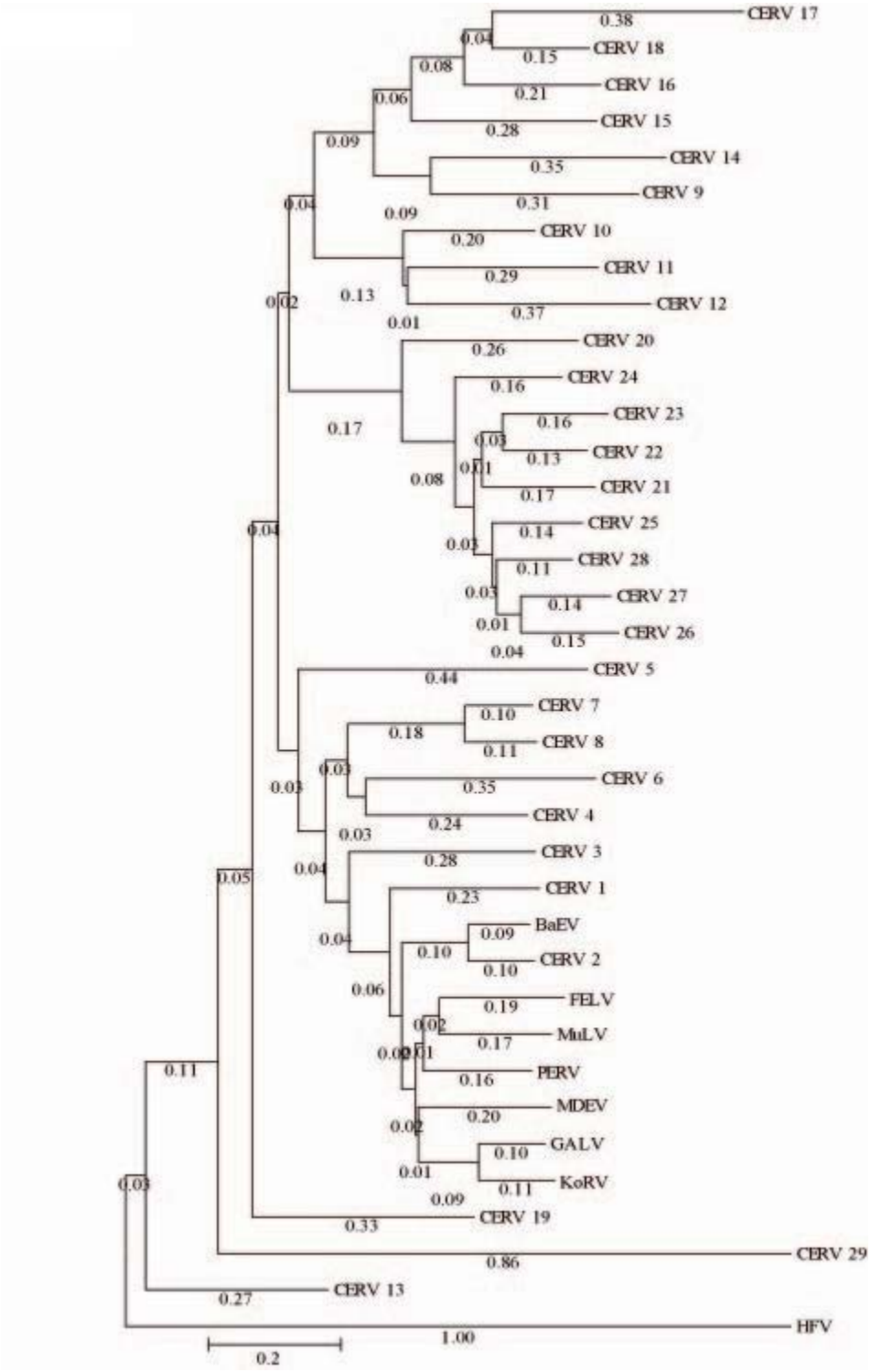

Figure 2.2: RT-based neighbor-joining tree for Class I chimpanzee endogenous retroviruses
The distances (corrected 'p' using Jukes-Cantor model) appear next to each of the branches. RT sequences from species other than chimpanzee, listed in table 2.1, are included for comparison.The outgroup is Class III element HFV (Human foamy virus; see Table 2.1 and Figure 2.5)

Because the long terminal repeats of endogenous retroviruses are synthesized from a single template during reverse transcription, they are identical at the DNA sequence level upon integration (Boeke and Stoye 1997). Using the primate pseudogene nucleotide substitution rate of 0.16 % divergence / million years (Kapitonov and Jurka 1996; Costas and Naveira 2000), the relative integration time or age of CERV elements can be estimated from the level of sequence divergence existing between the element's 5' and 3' LTRs. The Jukes-Cantor model was used to correct for the presence of multiple mutations at the same site, back mutations and convergent substitutions (Jukes and Cantor 1969) . Although caution must be taken when using LTR divergence to estimate the age of individual elements because of confounding processes such as recombination and conversion, [e.g., (Johnson and Coffin 1999; Hughes and Coffin 2005)], the method is able to provide useful age estimates, at least to a first approximation [e.g., (Bowen and McDonald 2001)]. Using this method, we estimate that the age of full-length Class I CERV elements range from 0.8 to 82.9 MY.

Full length elements representing at least three Class I CERV families, CERV 1/PTERV1, CERV 2 and CERV 3 (HERVS71) have been recently transpositionally active as evidenced by the presence of an unoccupied preintegration site at the corresponding locus in humans. Inconsistent with this view is the fact that one of the chimpanzee-specific CERV 3 (HERVS71) insertions located on the Y-chromosome displays an atypically high level of LTR-LTR sequence

divergence (9 %) indicative of it having inserted ~28 MYA. However, the clear absence of this insert both in the sequenced human genome (pre-integration site in tact) and in the genomes of several randomly sampled ethnically and geographically diverse humans (data not shown), indicates that this element most likely inserted after the chimpanzee-human divergence (~6 MYA) and that the exceptionally high level of LTR-LTR sequence divergence is due to an inter-element recombination or conversion event (Johnson and Coffin 1999; Hughes and Coffin 2005). All other Class I CERV elements are much older and have not been reproductively active since well before chimpanzees and humans diverged from a common ancestor.

Class II (families 30 - 39): The CERV families 30 through 39 group with Class II retroviruses (Figure 2.1, Figure 2.3). All Class II CERV families have orthologues in humans. The average size of full-length Class II CERVs is 7670 bp. This class of CERV elements range in size from 2,564 – 12,803 bp in length. As with Class I elements, much of the size variation among Class II elements is due to INDELs associated with non-functional elements. The average size of LTRs associated with full-length Class II CERV elements is 544 bp (range 243 – 1139 bp). Consistent with the fact that Class II CERVs are orthologous to human HERV K elements, all but one family of Class II CERV elements have lysine tRNA binding sites. The sole exception, CERV 39 (HERV K22), has a methionine tRNA binding site (Table 2.2). It has recently been proposed that HERV K22 be renamed HERV M to reflect its distinct primer binding site (Lavie et al. 2004). Unlike the other Class II CERV elements, the CERV 39 (HERV K22) family clusters closely with the betaretrovirus (MMTV, SRV-1) (Figure 2.1, Figure 2.3).

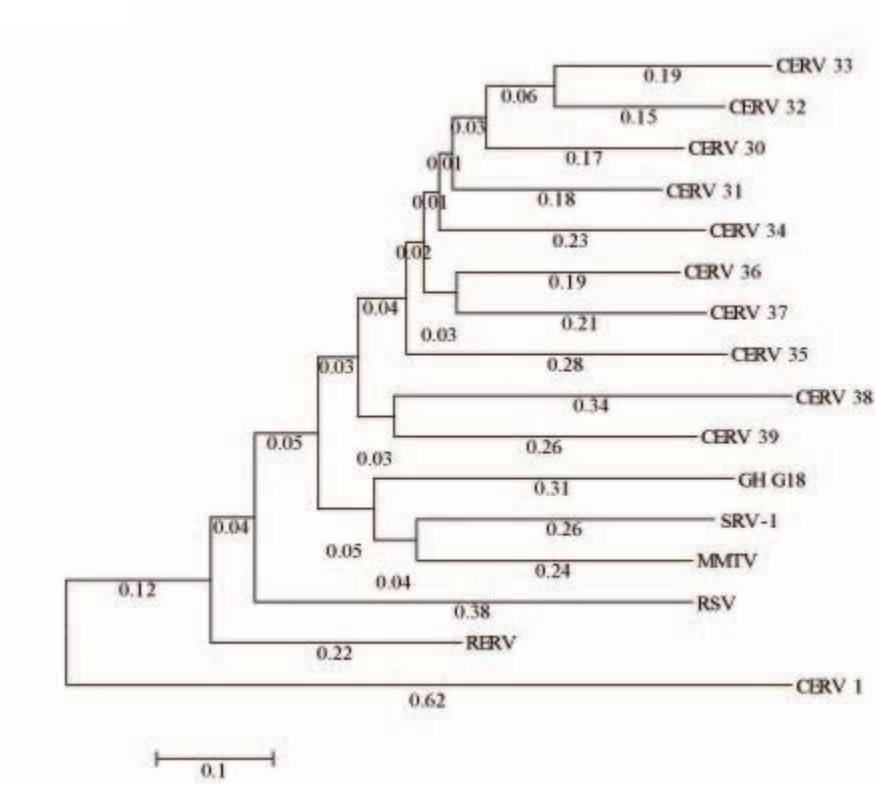

Figure 2.3: RT-based neighbor-joining tree for Class II chimpanzee endogenous retroviruses
The distances (corrected 'p' using Jukes-Cantor model) appear next to each of the branches. RT sequences from species other than chimpanzee, listed in table 2.1, are included for comparison. The outgroup is Class I element from CERV 1/PTERV1 family (see Table 2.2 and Figure 2.2)

The estimated age of full-length Class II CERV elements ranges from 2 to 97 MY. A member of only one Class II family, CERV 30 (HERV K10), has been transpositionally active since the divergence of chimps and humans from a common ancestor. The LTR sequence identity of one of the identified CERV 30 (HERVK10) elements is 99.4 %, indicating that this element inserted into the chimpanzee genome ~ 2 MYA. We have verified that this CERV 30 (HERV K10) insertion is absent in humans (Figure 2.4). It has been previously reported (Medstrand and Mager 1998; Barbulescu et al. 1999) and we found in our INDEL analysis (see below) that at least eight full length copies of CERV 30 orthologue HERV K10, inserted into the human

genome after the divergence of chimpanzees and humans from a common ancestor. In addition, two CERV 30 (HERV K10) insertion polymorphisms have been identified in human populations (Turner et al. 2001). Thus, CERV 30 (HERV K10) family members and their human orthologues have been transpositionally active in both human and chimpanzee lineages since these species diverged from a common ancestor ~ 6 MYA.

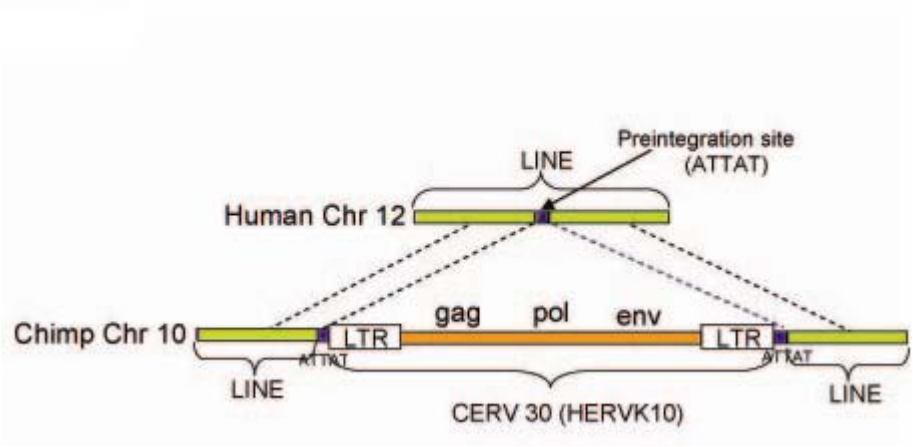

Figure 2.4 - Insertion of a member of CERV 30 (HERVK10) family in chimpanzees
The insertion occurred in the LINE element present in chromosome 10 of chimpanzee genome. The orthologous LINE element is present in chromosome 12 in humans. In chimpanzees target site duplications (ATTAT) are identified. A single copy of TSD (ATTAT, the preintegration site) is found inside the LINE element in humans. The LTRs of the element are 99.4% identical.

CERV 36 (HERV K11D) is the second oldest family of Class II CERV elements. We estimate that CERV 36 (HERV K11D) elements have not been transpositionally active for ~ 25 MY. We found that several members of the CERV 36 (HERV K11D) display the same deletion within the gag-pol regions of their genomes suggesting that this deletion occurred prior to their transposition. Thus, this sub-family of CERV 36 (HERV K11D) elements were, at one time, non-autonomous elements and acquired essential replicative functions in *trans*.

Class III (families 40 – 42): The CERV families 40 (HERV S), 41 (HERV 16) and 42 (HERV L) group with Class III retroviruses and are related to spumaviruses (Boeke and Stoye 1997) (Figure 2.1, Figure 2.5). All Class III CERV families have orthologues in humans. The average size of full-length Class III CERVs is 6758 bp. This class of CERV elements range in size from 2980 – 13271 bp in length. Again, much of this size variation is due to INDELs in this uniformly non-functional class of CERV elements. The average size of LTRs associated with full-length Class III CERV elements is 446 bp (range 254 – 831 bp). CERV 40 elements have a serine tRNA binding site while CERV 42 elements have a leucine tRNA binding site (Table 2.2). Due to sequence ambiguities, we were unable to determine the tRNA binding site for CERV 41 elements (Table 2.2). Class III CERV elements are the oldest group of endogenous retroviruses in the chimpanzee genome. The estimated age of these elements ranges from 30 to 145 MY.

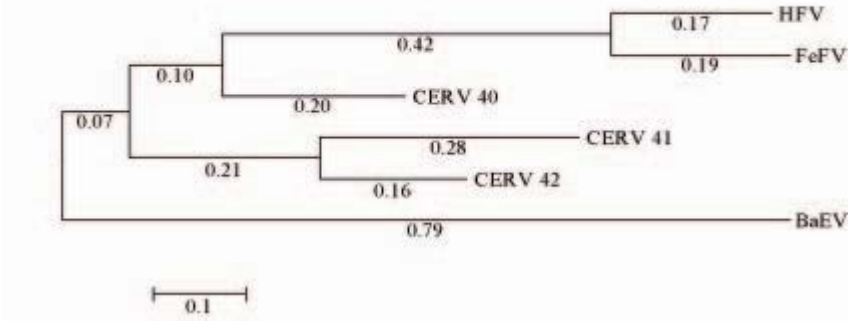

Figure 2.5: RT-based neighbor-joining tree for Class III chimpanzee endogenous retroviruses

The distances (corrected 'p' using Jukes-Cantor model) appear next to each of the branches. RT sequences from species other than chimpanzee, listed in table 2.1, are included for comparison. The outgroup is Class I element BaEV (Baboon endogenous retrovirus; see Table 2.1 and Figure 2.2)

Two CERV families have no human orthologues

CERV 1/PTERV1: With > 100 members, the CERV 1/PTERV1 is one of the most abundant families of endogenous retroviruses in the chimpanzee genome. CERV 1/PTERV1 elements range in size from 5 to 8.8 kb in length, are bordered by inverted terminal repeats (TG and CA) and are characterized by 4 bp TSDs (Table 2.2). The LTRs of CERV 1/PTERV1 family of elements range from 379 to 414 bp in length. CERV 1/PTERV1 elements have a proline tRNA primer binding site (Table 2.2). LTR sequence identity among CERV 1/PTERV1 elements ranges from 97.1 % to 99.7%.

Phylogenetic analysis of the LTRs from full length elements of CERV 1/PTERV1 members indicated that this family of LTRs can be grouped into at least two sub-families (bootstrap value of 99) (Figure 2.6). The age of each subfamily was estimated by calculating the average of the pairwise distances between all sequences in a given subfamily. The estimated ages of the two sub-families are 5 MY and 7.8 MY respectively suggesting that at least one sub-family was present in the lineage prior to the time chimpanzees and humans diverged from a common ancestor (~ 6 MYA). This conclusion, however, is inconsistent with the fact that no CERV 1/PTERV1 orthologues were detected in the sequenced human genome. Moreover, we were able to detect pre-integration sites at those regions in the human genome orthologous to the CERV 1/PTERV1 insertion sites in chimpanzees effectively eliminating the possibility that the elements were once present in humans but subsequently excised. Consistent with our findings, the results of a previously published Southern hybridization survey indicated that sequences orthologous to CERV 1/PTERV1 elements are present in the African great apes and Old World monkeys but not in Asian apes or humans (Yohn et al. 2005). These results suggest that some

members of the CERV 1/PTERV1 subfamily entered the chimpanzee genome after the split from humans through exogenous infections from closely related species and subsequently increased in copy number by retrotransposition. The unexpectedly high level of LTR-LTR divergence could be due to variation accumulated during the viral transfer (Belshaw et al. 2004) or possibly due to an inter-element recombination or conversion events subsequent to integration. Similar results were obtained when only the solo LTRs or both solo LTRs and LTRs from full length elements were used in constructing the phylogenetic trees (Figure 2.7, 2.8).

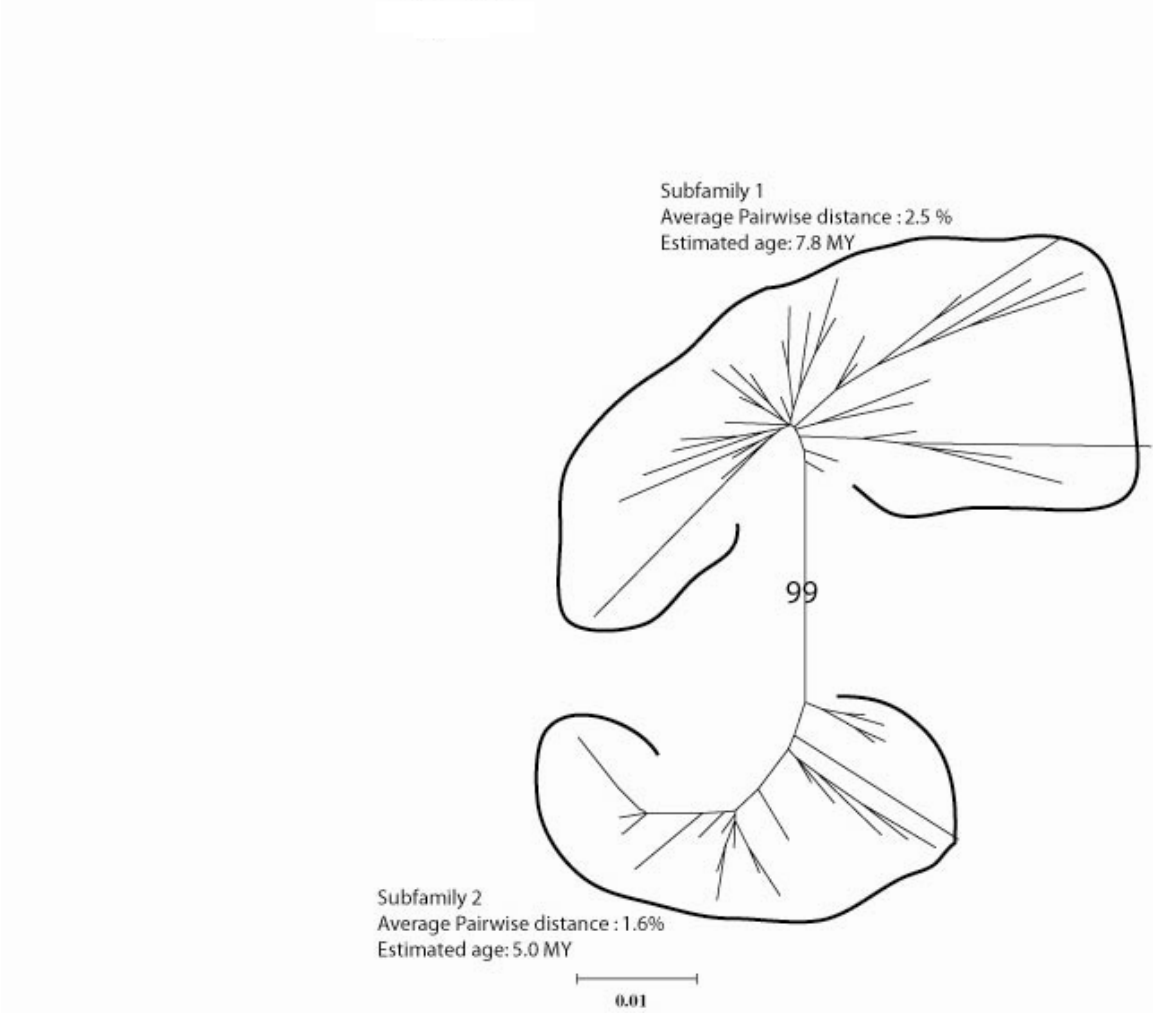

Figure 2.6 - Phylogenetic tree of CERV 1/PTERV1 LTRs

Unrooted neighbour joining phylogenetic tree built from 5' and 3' LTRs from full length CERV 1/PTERV1 elements. The average pairwise distances (corrected 'p' using Jukes-Cantor model) for each subfamily and the estimated ages are shown. Bootstrap values are shown.

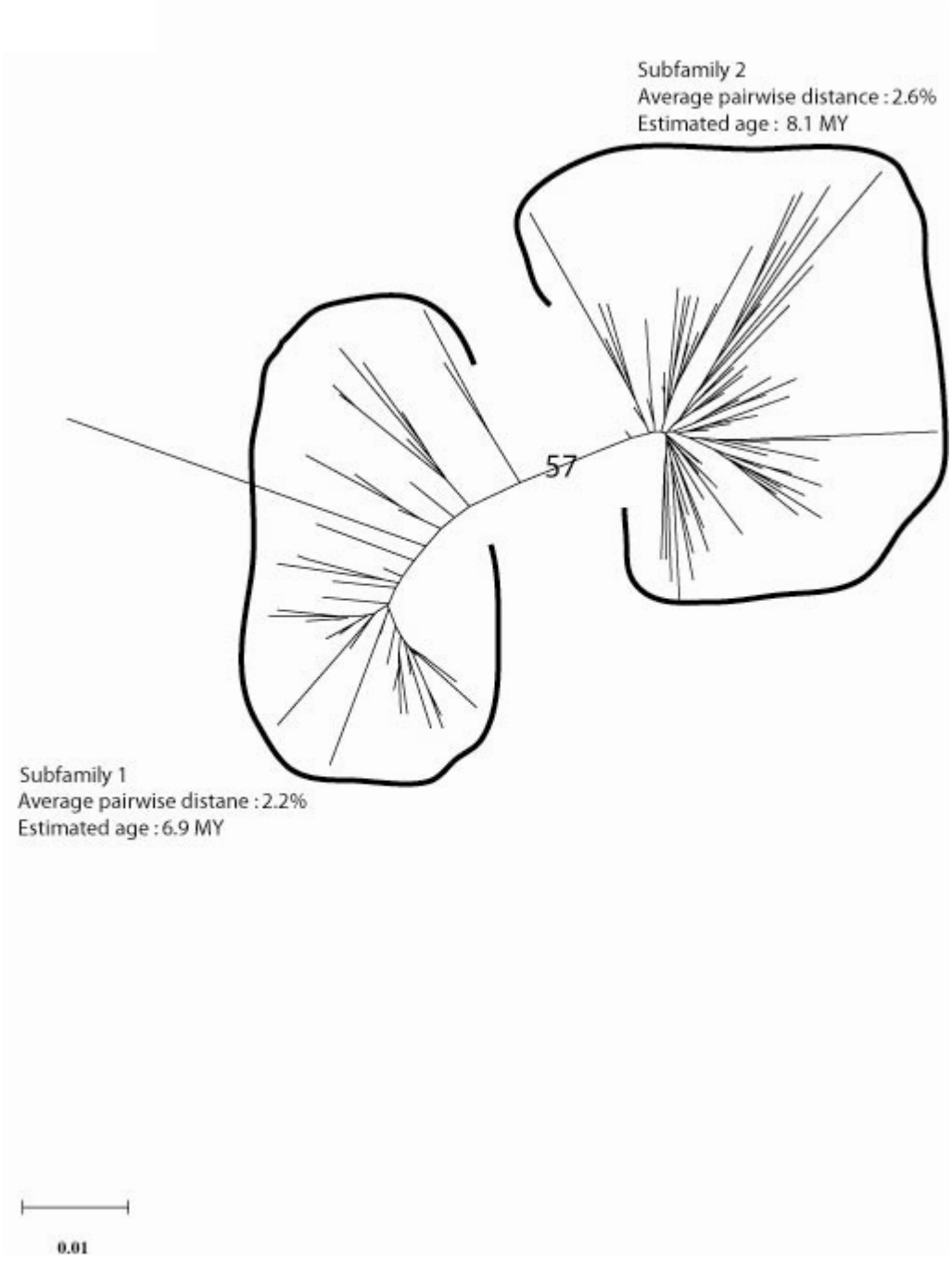

Figure 2.7: Unrooted neighbour joining phylogenetic tree built from solo LTRs and 5' and 3' LTRs of full length elements of CERV1/PTERV1 family

Bootstrap values are shown on the tree. The average pairwise distances (corrected 'p' using Jukes-Cantor model) for each subfamily and the estimated ages are shown.

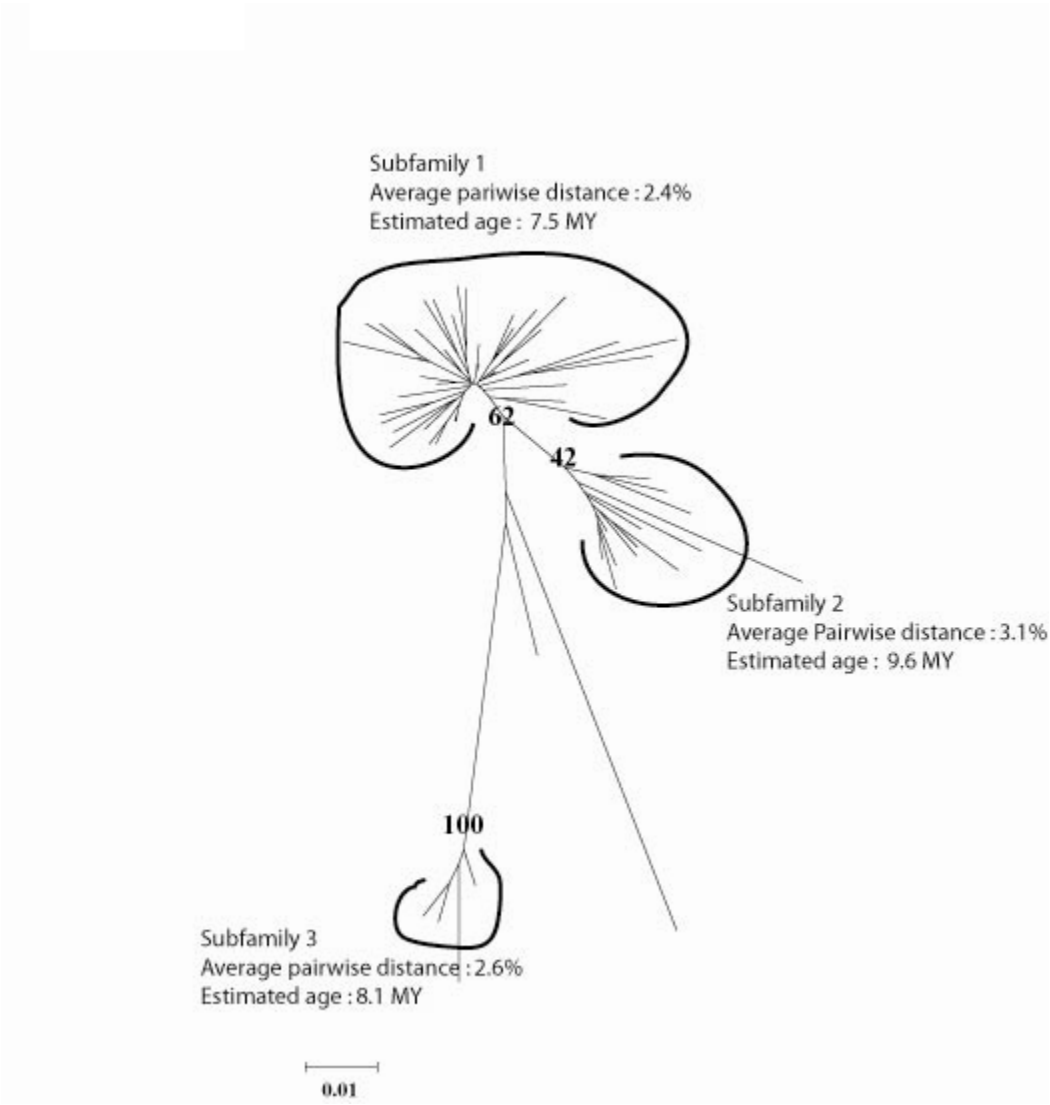

Figure 2.8: Unrooted neighbour joining phylogenetic tree built from solo LTRs of CERV 1/PTERV1 family
Bootstrap values are shown in the tree. The average pairwise distances (corrected 'p' using Jukes-Cantor model) for each subfamily and the estimated ages are shown.

We found that a number of CERV 1/PTERV1 elements with high (> 99%) LTR-LTR sequence identity have large (1-2 kb) deletions within the RT encoding region of their genomes. It is likely that these are non-autonomous elements that have inserted relatively recently by acquiring RT functions in *trans*, presumably from autonomous CERV 1/PTERV1 elements.

Instances of recently inserted LTR retrotransposons /endogenous retroviruses lacking RT-encoding functions have previously been detected in the genomes of humans (Smit 1993) and other species of both plants (Jiang et al. 2002; McCarthy et al. 2002) and animals [e.g., (Ganko et al. 2001)].

CERV 2: This is the second family of chimpanzee endogenous retroviruses with no orthologue in the human genome. We identified 10 solo LTRs and 8 full length copies of CERV 2 elements in the chimpanzee genome although, because of incomplete sequencing, we could identify the LTRs for only 4 of the 8 full length elements. CERV 2 elements are typically larger than CERV 1/PTERV1 elements, ranging in size from 8 – 10 kb in length. CERV 2 elements are bordered by inverted terminal repeats (TG and CA), have 4 bp TSDs (Table 2.2) and a proline tRNA primer binding site (Table 2.2).The LTRs of the CERV 2 family of elements range from 486 to 497 bp in length. Based on their LTR sequence identity (98.07 % to 99.6 %), we estimate that full length CERV 2 elements were transpositionally active in the chimpanzee genome between 1.3 – 6.0 MYA. Thus, the majority of CERV 2 elements were biologically active after the divergence of chimpanzees and humans from a common ancestor.

Phylogenetic analysis of solo LTRs and LTRs from full length elements revealed that CERV 2 elements group into at least four sub-families (bootstrap values > 95) (Figure 2.9). We estimated the ages of two of the more abundant subfamilies by calculating the average of the pairwise distances between all sequences in each sub-family. The estimated ages of the two sub-families were 21.9 MY and 14.1 MY respectively. As was the case for the CERV 1/PTERV1 family, these age estimates are inconsistent with the fact that no CERV 2 orthologues were detected in the sequenced human genome. Again, we were able to detect pre-

integration sites at those regions in the human genome orthologous to the CERV 2 insertion sites in chimpanzees effectively eliminating the possibility that the elements were once present in humans but subsequently excised.

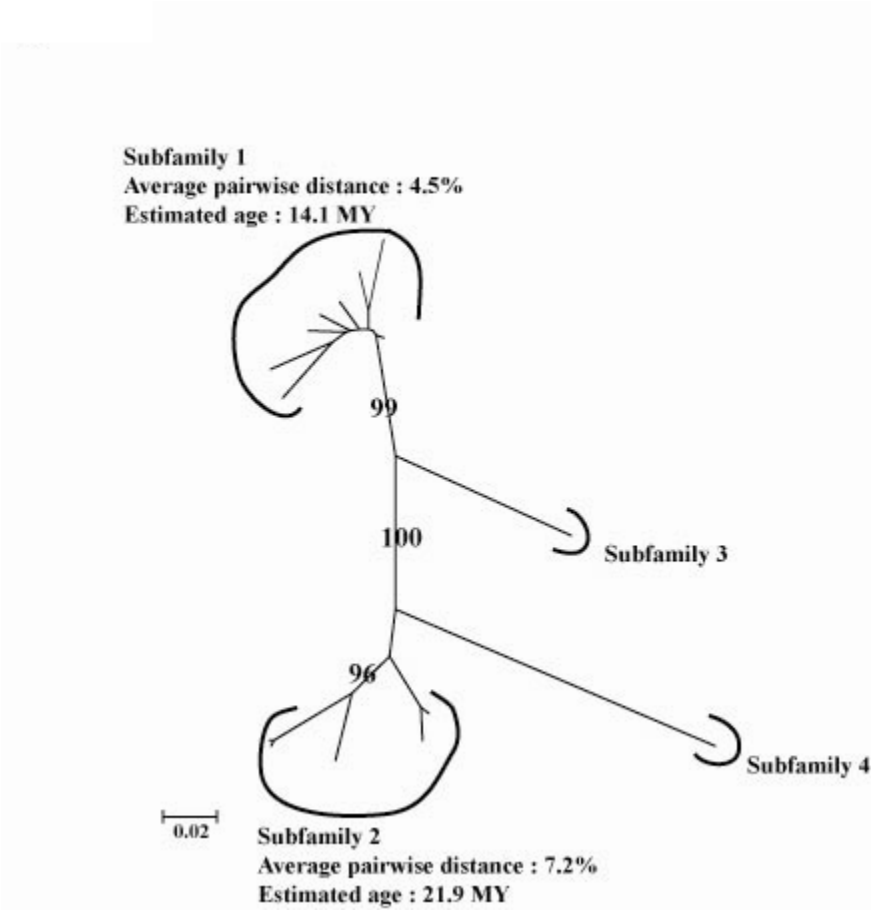

Figure 2.9 - Phylogenetic tree of CERV 2 LTRs
Unrooted neighbour joining phylogenetic tree built from CERV 2 solo LTRs and 5' and 3' LTRs from full length elements. The average pairwise distances (corrected 'p' using Jukes-Cantor model) for each subfamily and the estimated ages are shown. Bootstrap values are shown.

We assessed the distribution of CERV 2 elements in primates by PCR using primers complementary to sequences in the conserved RT region. The results indicate that CERV 2 elements are present in chimpanzee, bonobo and gorilla but absent in human, orangutan, old world monkeys, new world monkeys and prosimians (Figure 2.10a). Southern hybridization

experiments were carried out on DNA from species that gave negative PCR results to eliminate the possibility that the PCR primer binding sites have diverged in distantly related species within the CERV 2 RT and gag regions complementary to the designed probes (Figure 2.10b). The combined PCR and southern analysis indicate that CERV 2 like sequences are present in chimpanzee, bonobo, gorilla and old world monkeys but absent in human, orangutan, new world monkeys and prosimians (Figure 2.10c). This distribution of CERV 2 elements among primates is identical to the above described distribution of CERV 1/PTERV1 elements (Yohn et al. 2005). It is worth noting that although the probes used in southern hybridization were designed from chimpanzee element sequence, the strength of hybridization is higher in old world monkeys than in chimpanzees (Figure 2.10b) suggesting a higher copy number of CERV 2 elements in old world monkeys than in chimpanzees.

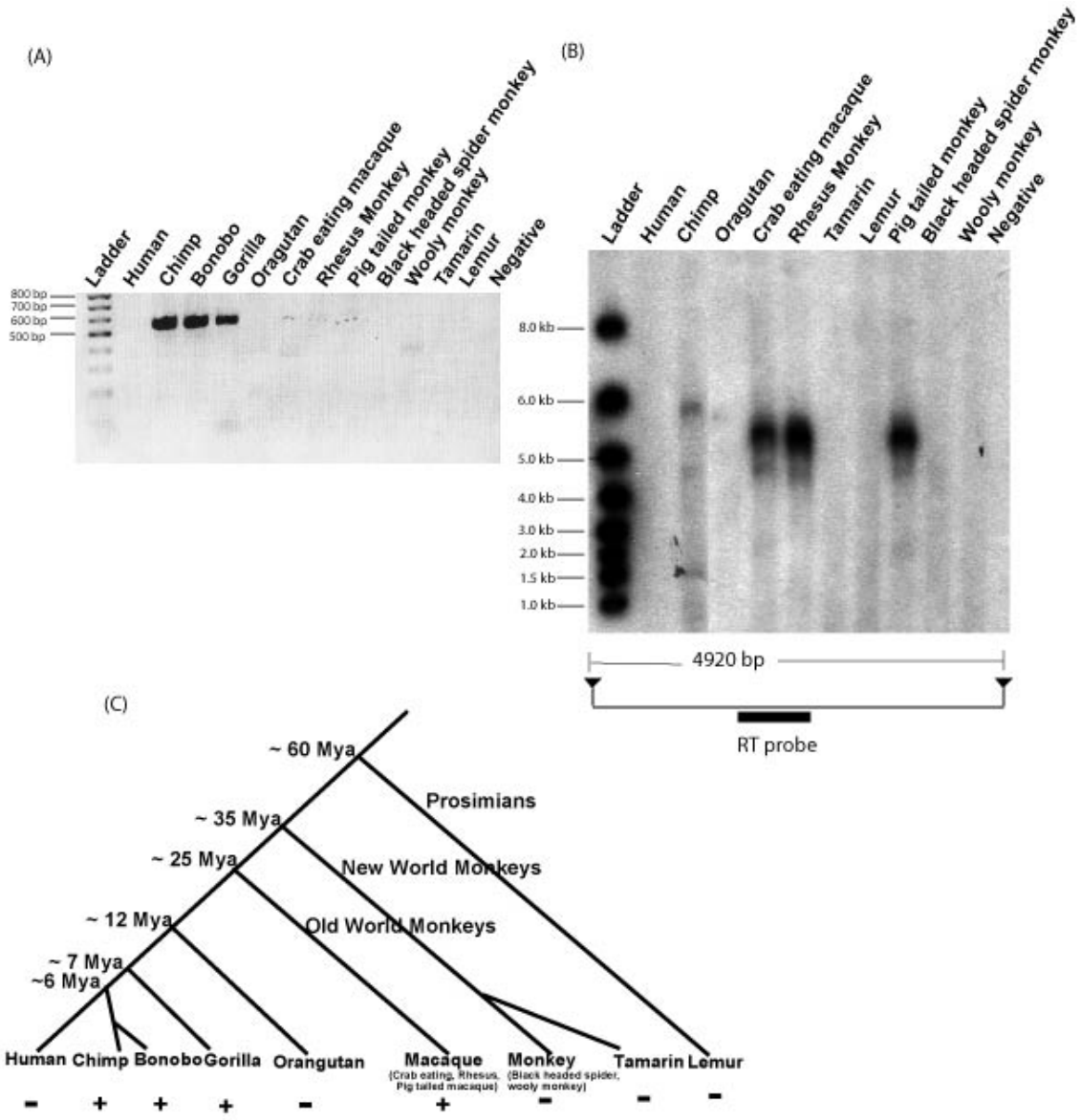

Figure 2.10 - Distribution of CERV 2 elements among primates

Species surveyed include Human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), crab eating monkey (*Macaca fascicularis*), rhesus monkey (*Macaca mulatto*, pig tailed monkey (*Macaca nemestrina*), black headed spider monkey (*Ateles geoffroyi*), wooly monkey (*Lagothrix lagotricha*), red-chested mustached tamari (*Saguinus labiatus*), ring-tailed lemur (*Lemur catta*).

a) PCR is conducted using primers designed in the RT region of chimpanzee CERV 2 element. The PCR results indicate that the CERV 2 element is present in chimpanzee, bonobo, gorilla and absent in other primates.

b) Southern hybridization is carried out on the DNA of the primates with negative PCR results using probe designed in the RT region. The results indicate that CERV 2 like elements are present in chimpanzee, crab eating macaque, Rhesus monkey and pig tailed

monkey. Though the same amount of DNA is loaded in all lanes, the strength of hybridization is higher in old world monkeys than in chimpanzees suggesting a higher copy number of CERV 2 elements in old world monkeys than in chimpanzees. Below the figure, a restriction map [chimpanzee sequence from chromosome 5 position 53871447.. 53880194 (NCBI Build 1 Version 1)] is presented in relation to the hybridization probe, HindIII (triangles)
c) The results from the combined PCR and southern analyses demonstrate a patchy distribution of CERV 2 elements among primates.

Endogenous retroviral positional variation between chimpanzees and humans

Comparative analyses of orthologous regions of the human and chimpanzee genomes has revealed a number of instances where relatively large spans of sequence present in one species are not present in the other (Britten 2002; Mikkelsen et al. 2005). It has been proposed that these gaps or INDELs may be of evolutionary significance [e.g., (Britten 1996)]. To determine the proportion of these gaps (Human gaps are sequences present in chimpanzees but absent in humans; Chimpanzee gaps are sequences present in humans but absent in chimpanzees) involving endogenous retroviruses, we utilized the human gap and chimpanzee gap datasets available at the UCSC Genome Bioinformatics web site (http://genome.ucsc.edu) that were generated by aligning the chimpanzee genome build panTro1 with the human genome build HG16 (Karolchik et al. 2003; Karolchik et al. 2004). These datasets include gaps of sizes ranging from 80 bp to 12.0 kb. Gap sequences from the datasets > 5000 bp (300 sequences), the typical length of full-length LTR retrotransposons/retroviruses, were blasted against the NCBI non redundant protein database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein) using BlastX (Altschul et al. 1990). BLAST was used to identify species-specific full length endogenous retroviral insertions in humans and chimpanzees. A total of 41 chimpanzee gap sequences and 31 human gap sequences were found to have significant similarity ($e < 0.01$) with retroviral sequences.

The presence of an endogenous retroviral sequence in chimpanzees that is missing at an orthologous genomic position in humans can be due to a novel insertion in chimpanzees or deletion of the element in humans. Similarly, the presence of an endogenous retroviral sequence in humans that is missing at an orthologous genomic position in chimpanzees can be due to novel insertion in humans or due to deletion of the element in chimpanzees. Because endogenous retroviruses do not precisely excise from insertion sites (Boeke and Stoye 1997), it is possible to distinguish between these two possibilities. If a region in humans orthologous to the position of an endogenous retroviral insertion in chimpanzees contains a remnant of endogenous retroviral sequence (e.g., fragmented element or solo LTR), we score the gap as a deletion in humans. If the orthologous region contains no remnant of the endogenous retrovirus but the pre-integration genomic sequence can be clearly identified, we score the gap as an insertion in chimpanzees. The same rules apply for the analogous dataset of the endogenous retroviral sequences present in humans but absent in chimpanzees.

Of the 41 instances where an endogenous retroviral sequence is present in chimpanzees but lacking in humans, 29 were due to novel insertions in chimpanzees while 12 were deletions in humans (Table 2.3, Figure 2.11). Of the 31 instances where an endogenous retrovirus is present in humans but absent in chimpanzees, we found that 8 were due to novel insertions in humans while 23 were deletions in chimpanzees (Table 2.4, Figure 2.12). Of the 29 novel insertions in chimpanzees 25 belong to CERV 1/ PTERV1 family, 2 to CERV 2 family, 1 to CERV 3 (HERVS7 1) family and 1 to CERV 30 (HERVK10) family whereas all the 8 novel insertions in humans belong to CERV 30 (HERVK10) family (Tables 2.3, 2.4). Thus, four families of

endogenous retroviruses have been transpositionally active in the chimpanzee lineage resulting in full length insertions since chimpanzees and humans diverged from a common ancestor while only one of these families [CERV 30 (HERVK10)] has been active in humans (Tables 2.3, 2.4). However, the family that is active in both humans and chimpanzees [CERV 30 (HERVK10)] generated 8 novel full length insertions in humans as opposed to only one novel insertion in chimpanzees since they diverged from the common ancestor (Tables 2.3, 2.4).

Table 2.3: Endogenous retrovirus INDEL sequences (> 5000 bp) present in chimpanzees but absent in humans

I: Insertion in chimpanzees, D: Deletion in humans

Human gaps	**Gap sequence**
25 (I)	CERV 1/PTERV1
2 (I)	CERV 2
2 (D)	CERV 18 (HERV9)
2 (D)	CERV 32 (HERVK14C)
2 (1 I + 1 D)	CERV 30 (HERVK10)
2 (D)	CERV 42 (HERVL)
1 (D)	CERV 17 (HERV30)
1 (I)	CERV 3 (HERVS71)
1 (D)	CERV 28 (HERVIP10F)
1 (D)	CERV 7 (Harlequin)
1 (D)	LTR1D
1 (D)	CERV 34 (HERVK9)

Table 2.4: Endogenous retrovirus INDEL sequences (> 5000 bp) present in humans but absent in chimpanzees
I: Insertion in humans, D: Deletion in chimpanzees

Chimpanzee gaps	Gap sequence
9 (8 I +1 D)	CERV 30 (HERVK10)
10 (D)	CERV 11 (HERVH)
2 (D)	CERV 18 (HERV9)
1(D)	CERV 16 (HERV17)
1 (D)	CERV 34 (HERVK9)
1 (D)	CERV 37 (HERVK11)
1 (D)	CERV 35 (HERVK13)
1 (D)	CERV 3 (HERVS71)
1 (D)	CERV 42 (HERVL)
2 (D)	CERV 7 (Harlequin)
1 (D)	CERV 28 (HERVIP10F)
1 (D)	CERV 19 (PABL_B)

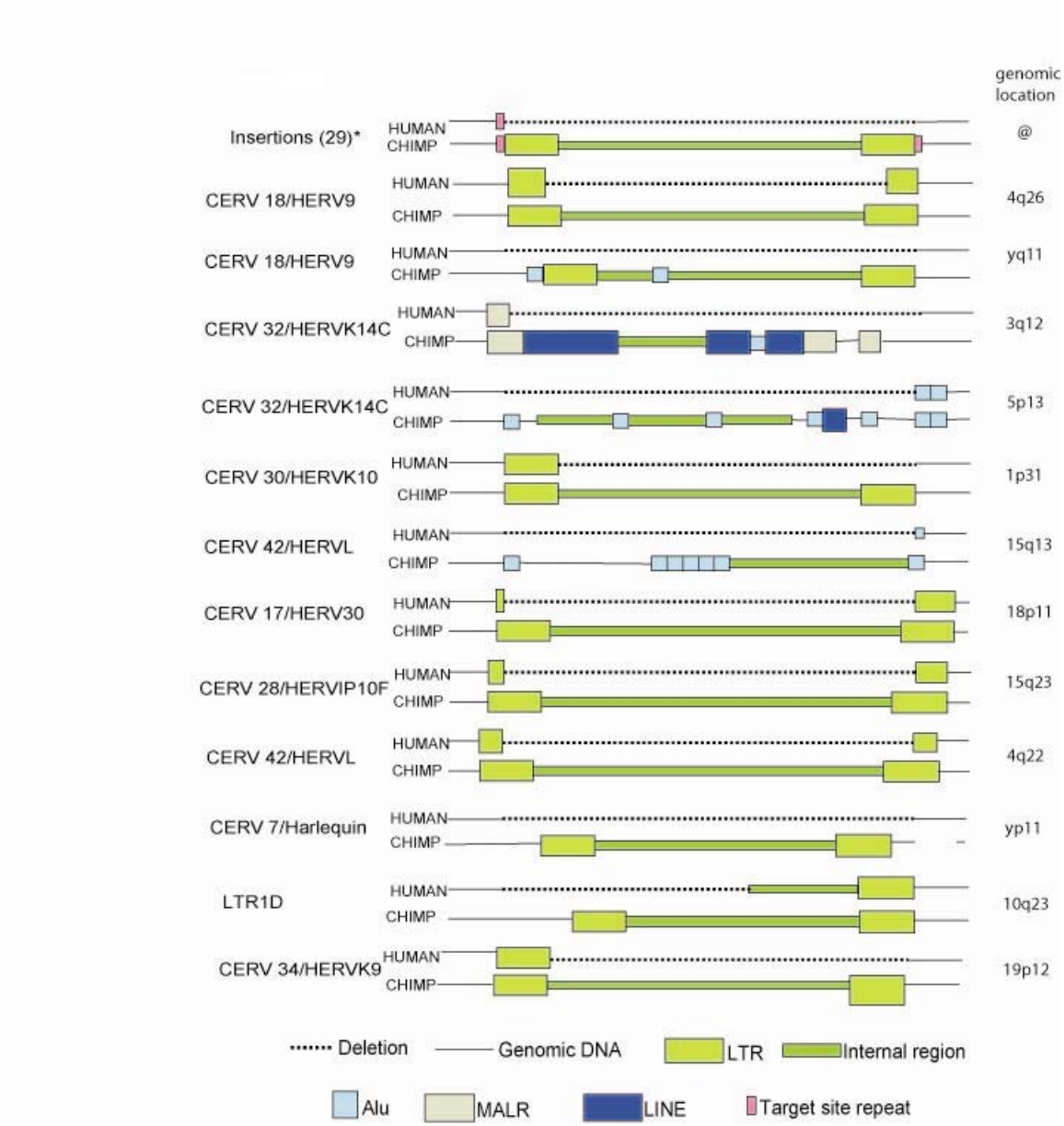

Figure 2.11 - Structure of endogenous retroviral INDEL sequences (> 5000 bp) in humans

Characteristics of the remnant endogenous retroviral sequences (solo LTRs and/or fragmented elements) in humans. [* indicates 29 chimpanzee specific endogenous retroviral insertions of which 25 belong to CERV 1/PTERV1 family [@ genomic locations: 5p12,1q25,12p13,10q11,3q11,14q23, 2p11,12p11,4q13,5q11,3p22,11p12,13q12,6q25,7p14,16p13,7p21,3q12,12q12,4p12,5p14 ,10q21,7p14, 7p11,5q12], 2 belong to family CERV 2 [@ genomic locations: 3p24,19q13], 1 to CERV 3 (HERVS71) family [@ genomic location: yp11] and 1 to CERV 30 (HERVK10) family [@ genomic location: 12p13] (see table 2.3)]

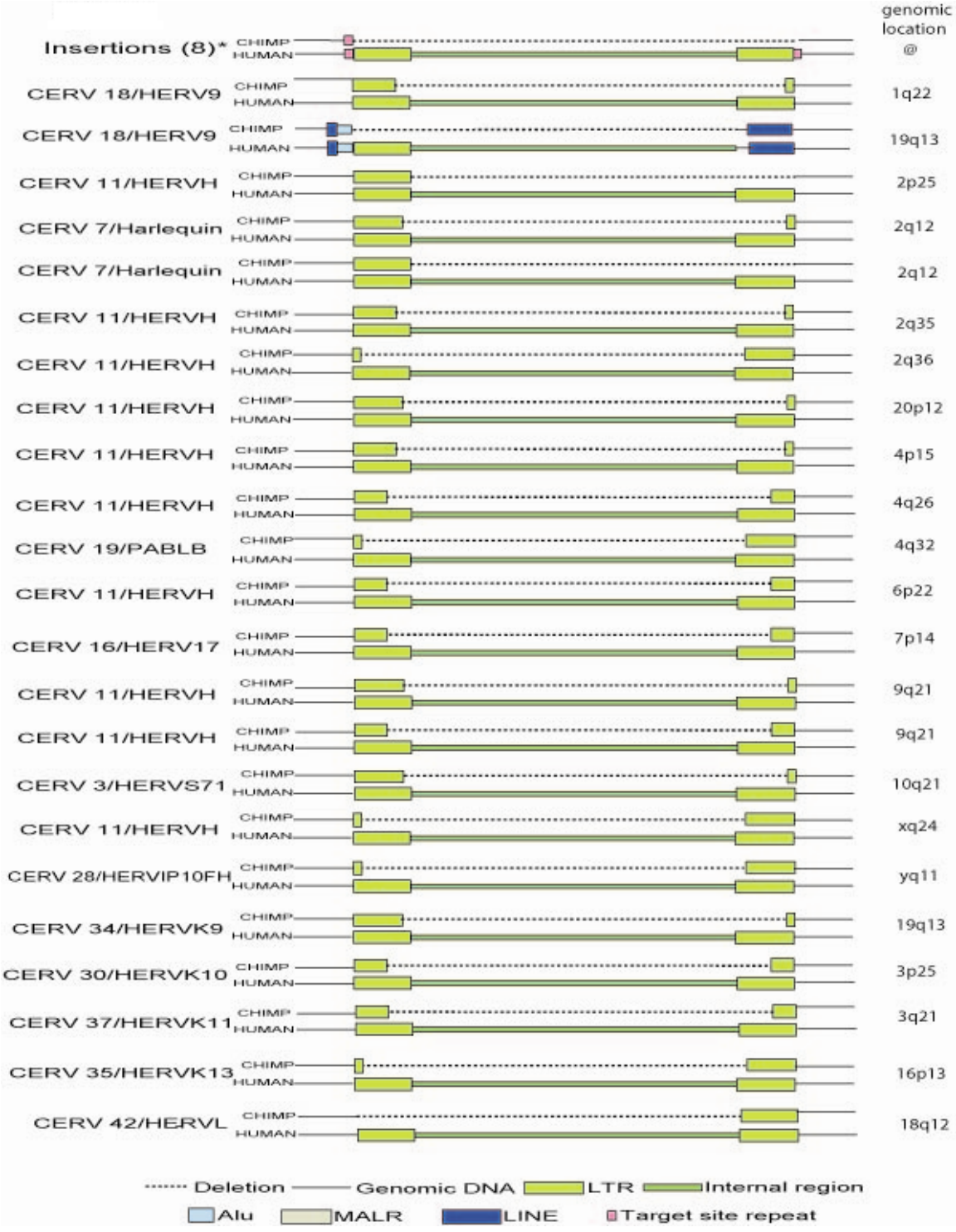

Figure 2.12 - Structure of endogenous retroviral INDEL sequences (> 5000 bp) in chimpanzees

Characteristics of the remnant endogenous retroviral sequences (solo LTRs and/or fragmented elements) in chimpanzees.[* indicates 8 human specific endogenous

retroviral insertions from CERV 30 (HERVK10) family [@ genomic location: 22q11, 3q13, 3q27, 5q33, 6q14, 10q24, 11q22, 12q14] (see table 2.4)]

Since solo LTRs and fragmented endogenous retroviral copies are typically ten to a hundred times more abundant than full length elements in humans (Stoye 2001; Polavarapu et al. 2006a) we extended our survey to determine the extent to which INDEL variation between humans and chimpanzees is associated with solo LTRs and/or fragmented endogenous retroviral sequences. We again utilized datasets (Human gaps and Chimpanzee gaps) available at UCSC Genome Bioinformatics web site (http://genome.ucsc.edu). We used "Repeat Masker" (A.F. Smit and P. Green, unpublished data) to identify all interspersed repeats i.e. all transposable elements present in the datasets and to subsequently extract endogenous retroviral homologous sequences.

Gap sequences were divided into two types: 1) "Mosaic type" gap sequences are defined as those comprised of more than one category of interspersed repeats. (e.g., endogenous retrovirus inserted within a LINE element); and 2) "Single type" gap sequences are defined as those comprised of only sequences homologous to endogenous retroviruses. Single type gap sequences were further divided into two categories: a) Category 1 are those gap sequences comprised entirely of an endogenous retroviral sequence while b) Category 2 are those gap sequences comprised of endogenous retrovirus and non-interspersed repeat sequences. The above categorizations are useful in distinguishing gaps due to deletions in one species from the gaps due to insertions in the other species. Instances of mosaic type and single type category 2 gaps are deletions in that species while the gaps that belong to single type category 1 are either deletions in that species or

insertions in the other species. Because endogenous retroviruses do not excise precisely (Boeke and Stoye 1997) from the insertion sites these later gaps can be further characterized as the result of insertions or deletions.

We found a total of 18,395 human gap sequences of which 9,855 (53.57 %) contained interspersed repeats. Chimpanzees had a total of 27,728 gap sequences of which 15,652 (56.44 %) contained interspersed repeats. A total of 1495 human gap sequences contained endogenous retroviral sequences [592 mosaic type and 903 single type (category 1: 640 + category 2: 263)] (Table 2.5). Five hundred and ninety two mosaic types and 263 single type category 2 were deletions in humans. Of the 640 single type gaps belonging to category 1, 151 are new insertions in chimpanzees while the remaining 489 are deletions in humans. Of the 151 chimpanzee insertions, 97 involved the two chimpanzee specific families while the remaining involved CERV families with human orthologues (Table 2.5).

A total of 1608 chimpanzee gap sequences contained endogenous retroviral sequences [758 mosaic type and 850 single type (category1: 557 + category 2: 293]. As stated above, 758 mosaic types and 293 single type category 2 are deletions in chimpanzees. Of the 557 that belonged to single type category 1, 79 are new insertions in humans while the remaining 478 are deletions in chimpanzees (Table 2.5).

Table 2.5: Endogenous retrovirus INDELs (80 bp - 12 kb) in humans and chimpanzees
(* indicates solo LTRs and/or fragmented copies)

	Humans	Chimpanzees
Total Gaps	18,395	27,728
Gaps with interspersed repeats	9,855	15,652
Gaps containing endogenous retrovirus sequences	1495	1608
Mosaic gaps	592	758
Single gaps	903	850
Category 2 gaps	263	293
Category 1 gaps	640	557
Deletions	489	478
Insertions	79	151
CERV 1/PTERV1	-	85 (25 full ln + 60 solo LTRs)
CERV 2	-	12 (2 full ln + 10 solo LTRs)
CERV 30 (HERVK10)	78 (8 full ln + 70 solo LTRs)	43 (1 full ln + 42 solo LTRs)
CERV 3 (HERVS71)	-	1 (1 full ln)
CERV 11 (HERVH/LTR7)	-	1*
CERV 37 (HERVK11/MER11C)	-	1*
CERV 34 (HERVK9/MER9)	-	1*
CERV 18 (HERV9)	-	7*
CERV 35 (HERVK13/LTR13)	-	1*
MER31B	1*	-

Consistent with the copy number of CERV 30 (HERVK10) species-specific full length insertions (Tables 2.3, 2.4), the insertions of solo LTRs from this family were higher in humans (78) than in chimpanzees (43) since they diverged from the common ancestor

(Table 2.5). Apart from CERV 30 (HERVK10) insertions in both the genomes, five other endogenous retroviral families continued to be active in chimpanzees resulting in solo LTR or fragmented insertions while only one new insertion from MER31B occurred in humans since they diverged from the common ancestor (Table 2.5).

CONCLUSIONS

Once considered parasitic sequences of little or no adaptive significance (Doolittle and Sapienza 1980; Orgel and Crick 1980), transposable elements are today generally recognized as significant contributors to human regulatory [e.g., (Jordan et al. 2003)] and structural [e.g., (Nekrutenko and Li 2001)] gene evolution. The recent sequencing of the chimpanzee genome is providing a unique opportunity to conduct comparative genomic analyses of primate transposable elements.

Retrotransposons are the most abundant class of transposable elements. For example, retrotransposons comprise at least 60% of the human genome (Lander et al. 2001) and results presented here and elsewhere (Mikkelsen et al. 2005) suggest that the number of endogenous retroviruses in chimpanzees may be higher than in humans. In this paper, we present the results of the first systematic search for endogenous retroviruses in the chimpanzee genome. We have identified 425 full-length endogenous retroviruses in the chimpanzee genome that can be grouped into 42 independent lineages or families (Figure 2.1). All but two families of chimpanzee endogenous retroviruses were found to have orthologues in humans (Table 2.2). In contrast, we have found that all known families of human endogenous retroviruses have

orthologues in chimpanzees. The two CERV families without orthologues in the human genome display a patchy distribution among primates (Figure 2.10) and our data suggest that at least some members of both families have been transpositionally active in the chimpanzee lineage after the divergence of chimpanzees and humans from a common ancestor.

The absence of elements from two CERV families (CERV 1/PTERV1 and CERV 2) in the sequenced human genome could be explained by either these elements present in the common ancestor of humans and chimpanzees and subsequently excised from the human genome or that these elements infected chimpanzee genome after the divergence of humans and chimpanzees from the common ancestor. The presence of pre-integration sites at those regions in the sequenced human genome orthologous to the CERV 1/PTERV1 and CERV 2 insertion sites in chimpanzees effectively eliminate the possibility that these elements were once present in the sequenced human genome but were subsequently excised. Rejecting the possibility that CERV 1/PTERV1 elements are present in the human population through identification of pre-integration sites in the sequenced human genome could however be challenged by the scenario that these elements are polymorphic in the human population. To address this scenario, we experimentally (PCR) checked for the presence of an oldest CERV 1/PTERV1 element from the chimpanzee genome (LTR-LTR sequence divergence of 4.06% aging to 13 MY indicating that this element is present in the common ancestor of humans and chimpanzees and would be the most likely candidate to check for polymorphism among humans if such an event existed) among 30 human samples from ethnically different races around the world. Results from PCR showed that this element is absent in world wide human populations indicating that elements from CERV 1/PTERV1 families are not polymorphic in human population (data not shown).

Another scenario explaining the patchy distribution of CERV 1/PTERV1 and CERV 2 elements may be that the primate phylogeny is not applicable to the whole genome and that for certain regions in the genome, humans and orangutans form a single clade distinct from chimpanzee-gorilla clade. This alternative primate phylogeny has been proposed by few anthropologists (Schwartz 1984). The alternative primate phylogeny seems unlikely in light of the extensive molecular evolutionary data that have been collected over the last few years (Goodman 1999; Chen and Li 2001) that clearly place orangutan as the outgroup species to the human–chimpanzee–gorilla clade and Old World monkeys as an outgroup to the human/ape lineage. Also, the absence CERV 1/PTERV1 elements at orthologous sites among chimpanzee, gorilla, macaque and baboon further eliminates the possibility of dissent from the common ancestor even with the alternate primate phylogeny (Yohn et al. 2005).

The lack of evidence supporting the presence of CERV 1/PTERV1 and CERV 2 elements in the chimpanzee genome through dissent from common ancestor indirectly indicate the alternative possibility that these two families arose in the chimpanzee genome by exogenous infection and subsequently increased in copy number through transpositional events. The unexpectedly high level of LTR-LTR divergence among members of CERV 1/PTERV1 and CERV 2 families could be due to variation accumulated during the viral transfer (Belshaw et al. 2004) or possibly due to an inter-element recombination or conversion events subsequent to integration. However, the source of infection is not known. One possibility may be that these elements were introduced independently and

contemporarily into african ape lineages by horizontal transmission, perhaps from contact with ancient old world monkeys while contemporary human and orangutan lineages escaped such infections. This scenario could be supported by the geographic isolation of the african and asian ape lineages during the Miocene period (Chaimanee et al. 2003; Kunimatsu et al. 2004). However, the presence of CERV 1/PTERV1 elements in both asian (macaque) and african (baboon) old world monkeys indicate that the exogenous source virus is endemic to both continents eliminating the above possibility of absence of elements in asian apes and humans due to geographic isolation. Also, the ancestral habitat of early hominids is generally thought to have overlapped, in part, with the african apes (WoldeGabriel et al. 2001; Brunet et al. 2002).

Therefore, the most likely scenario for the absence of CERV 1/PTERV1 and CERV 2 elements in humans and asian apes might be that humans and asian apes developed resistance to infections by these viruses whereas african apes were susceptibile to such infections. Evidence supporting this scenario has emerged recently indicating that such resistance mechanism does exist in humans (Kaiser et al. 2007). Selective changes have occurred in human lineage in immune protein TRIM5ά conferring resistance to humans for CERV 1/PTERV1 infections (Kaiser et al. 2007). It is also suggested that selective changes that have occurred in human lineage during acquisition of resistance to CERV 1/PTERV1 infection might have left humans more susceptible to infection by human immunodeficiency virus type 1 (HIV-1) (Kaiser et al. 2007). The presence of such resistance mechanism in orangutans would further support this scenario but has yet to be discovered.

We estimate that chimpanzee endogenous retroviruses range in age from ~0.8 to 145 MY. Nine families of chimpanzee endogenous retroviruses have been transpositionally active in chimpanzees while two families of human endogenous retroviruses have been transpositionally active in humans since they diverged from a common ancestor (Table 2.5). Thus, while some families of endogenous retroviruses have not been transpositionally active within the primate lineage, others have and continued to be active since chimpanzees and humans diverged from a common ancestor.

It has been estimated that 3.5% of the sequence differences between chimpanzees and humans is due to INDELs (Britten 2002; Mikkelsen et al. 2005) and that this INDEL variation may be of particular evolutionary significance (Britten 1996). We have determined that ~7% of all chimpanzee-human INDEL variation is attributable to the presence or absence of endogenous retroviral sequences. The potential biological/evolutionary significance of this variation is currently under investigation.

Emerging evidence indicates that retrotransposon have played a significant role in gene and genome evolution [e.g., (McDonald 1993; Britten 1996; Brosius 1999b)]. The identification, characterization and comparative genomics of chimpanzee endogenous retroviruses presented in this report should not only help contribute to our understanding of the functional significance of these elements in chimpanzees but to a better appreciation of the role of endogenous retroviruses in primate evolution.

MATERIALS AND METHODS

Initial dataset scanning

The 2.73 GB chimpanzee genomic sequence (http://www.ensembl.org/Pan_troglodytes/) obtained from the Ensembl database was scanned for the presence of endogenous retroviruses using a structure based program, LTR_STRUC (LTR retrotransposon structure program) (McCarthy and McDonald 2003). LTR_STRUC scans the genomic sequence for the presence of similar regions of length typical for LTRs (LTR pairs) and within the expected size of a full length LTR retrotransposon/endogenous retrovirus. If the putative LTRs are found, the program then searches for additional retrotransposon features like Primer Binding Site (PBS), Poly Purine Tract (PPT), Target Site Repeats (TSR) and assigns a reliability score to the hit based on presence or absence of each of these features. A total of 2056 hits were reported as the putative endogenous retroviruses in the chimpanzee genomic sequence, of which only 97 encoded Reverse Transcriptase.

Sequence analysis for identifying the RT coding sequence

The 97 putative elements for which the presence of RT sequence was reported by LTR_STRUC were subjected to sequence analysis to identify the Reverse Transcriptase (RT) coding region. Briefly, sequence analysis involves aligning the amino acid sequence of the three reading frames reported by the search algorithm (the strand encoding RT protein is determined based on the presence of PBS and PPT) with previously annotated retroviral proteins using ClustalX (Thompson et al. 1997) followed by manually checking the three ORFs for the RT conserved motifs previously described (Xiong and Eickbush 1988, 1990). From this sequence analysis we were able to identify RT conserved motifs for 25 hits.

Identification of additional elements

The 25 RT sequences obtained from sequence analysis were augmented by conducting exhaustive sequence similarity searches using these sequences as queries against the 2.73 GB chimpanzee genomic sequence (http://www.ensembl.org/Pan_troglodytes/) using TBLASTN program (Altschul et al. 1990; Altschul et al. 1997) to obtain an extensive set of endogenous retroviruses in the sequenced genome. Around 2000 RT sequences were obtained by automatically parsing the TBLASTN search results for the hits above a threshold of 70% identity and covering a length of one third of the query sequence using a perl script. After removing duplicates obtained during automatic parsing, we were left with 1088 RT sequences.

Identification of full length elements

The 1088 RT sequences identified in the TBLASTN searches were checked for the presence of LTRs on either side of the RT as a criterion for full length elements. This was done by examining the DNA sequences 7000 bp on either side of the RT sequence, aligning them against each other using BLAST2SEQ program and manually checking the hits for the presence of canonical dinucleotides, target site repeats and other LTR characteristic features. LTRs were identified for 395 of the 1088 elements that had RT sequences. Thirty elements from previously reported human RT sequences for which orthologues were not identified in the above searches were added to our dataset resulting in a total of 425 elements.

Multiple sequence alignments and phylogenetic analysis

A multiple alignment was constructed from the DNA sequences of the RT region of 425 full length elements together with representative members from the three classes of vertebrate retroviruses/ LTR retrotransposons (Table 2.1) (Boeke and Stoye 1997) using ClustalX program (Thompson et al. 1997). We chose to use DNA sequence in making the multiple alignment and building the phylogenetic tree rather than amino acid sequence because of the presence of numerous frame shift mutations and stop codons in the elements. The multiple alignment was manually adjusted in the MEGA alignment browser (Kumar et al. 2004). A neighbor joining tree was generated from the alignment using MEGA2 with p-distance and pairwise deletions as parameters and bootstrap values were obtained from 1000 replicates.

Grouping the elements into families

The full length elements were grouped into families based on the bootstrap values generated in the phylogenetic tree. Phylogenetically well supported clusters with high boot strap values were used to group the elements into families (Figure 2.1). The most recent element that is still intact is used as the representative element for each family (Table 2.2).

Identification of the primer binding site

A 100bp region downstream of the 5' LTR of full length elements was searched against the chimpanzee tRNA database downloaded from http://lowelab.ucsc.edu/GtRNAdb/Ptrog/ (Lowe and Eddy 1997) using FASTA program (Pearson and Lipman 1988). The 3' end of tRNA that matched with the reverse complement of the sequence over a stretch of 14-22 bp was assigned as a tRNA primer of the element (Table 2.2).

Evolutionary Analysis of CERV 1/PTERV1 and CERV 2 LTR sequences

Mulitple alignment was generated from the LTRs for each family using ClustalX (Thompson et al. 1997). A neighbor joining tree was generated from the alignment using MEGA3 (Kumar et al. 2004) with Jukes-Cantor model (Jukes and Cantor 1969) and pairwise deletions as parameters and bootstrap values were obtained from 1000 replicates. The age of each subfamily was estimated by calculating the average of pairwise distances between all sequences in that subfamily and using the primate pseudogene nucleotide substitution rate of 0.16 % divergence / million years (Kapitonov and Jurka 1996; Costas and Naveira 2000).

Molecular analysis (PCR and Southern Hybridization)

Primate DNA samples were purchased form coriell cell repository (catalog No# PRP00001 and PRP00003).

Polymeraze chain reaction: Primers were designed in the conserved RT, gag, LTR and env regions of the CERV 2 element using PRIMER3 program (Rozen and Skaletsky 2000). PCR amplification conditions were as follows. (Initial denaturation for 4 min and 30 sec at 94°C, 30 cycles of 30 sec denaturation at 94º C, 30 sec annealing at 57º C, 40 sec elongation at 72º C and a final 1-cycle extension of 7 min at 72º C. The PCR products were then visualized on 1% (w/v) agarose gel.

Southern Hybridization: Primate DNA was restriction enzyme digested, transferred to nylon membrane and hybridized as described previously (Daly et al. 2000). Nested PCR amplified products in RT and gag regions of CERV 2 elements were radioactively

labeled and used as probes for hybridization. The same amount of DNA was loaded in all the lanes. DNA samples in order: Human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus*), crab eating monkey (*Macaca fascicularis*), rhesus monkey (*Macaca mulatto*), red-chested mustached tamari (*Saguinus labiatus*), ring-tailed lemur (*Lemur catta*), pig tailed monkey (*Macaca nemestrina*), black headed spider monkey (*Ateles geoffroyi*), wooly monkey (*Lagothrix lagotricha*).

ACKNOWLEDGEMENTS

We thank Lilya Matyunina for help with Southern hybridization. This research was supported by a grant from the Georgia Institute of Technology Research Foundation.

CHAPTER 3

NEWLY IDENTIFIED FAMILIES OF HUMAN ENDOGENOUS RETROVIRUSES (HERVS)

INTRODUCTION

Human endogenous retroviruses (HERVs) comprise approximately 8.3% of the human genome (Lander et al. 2001). HERVs have previously been classified into 31 distinct families based upon sequence alignment of reverse transcriptase (RT) and envelope (ENV) domains and subsequent phylogenetic analyses (Jurka 2000; Tristem 2000; Benit et al. 2001). Using the data mining program, LTR_STRUC (McCarthy and McDonald 2003) in conjunction with conventional sequence homology techniques, we recently completed an analysis of chimpanzee long terminal repeat (LTR) retrotransposon families (in preparation). Since LTR_STRUC searches for LTR retrotransposons based on structure (e.g., presence of LTRs , target site duplications, tRNA binding sites, etc) rather than homology, elements can be identified that go undetected in traditional BLAST searches. We identified 9 chimpanzee endogenous retrovirus families that are orthologous to HERV families not previously identified. These 9 newly discovered HERV families are described and characterized in this letter.

RESULTS

LTR retrotransposons and retroviruses are grouped into three major classes (Griffiths 2001). Class I contains elements related to gammaretroviruses, Class II elements are related to betaretroviruses and Class III elements are distantly related to spumaviruses.

The RT based phylogeny indicates that all the newly identified HERVs described here are Class I elements (Figure 3). The detailed characteristics of each of the newly discovered HERV families are presented in the Table 3.1 and Table 3.2. All are low abundance families being comprised of only 1-7 full-length members with low homology to previously identified HERVs. This may, in part, explain why they have not been previously identified. The newly discovered full-length elements are of standard HERV length (7,198-10,675 bp with 359-682 bp LTRs) and display typically sized target site duplications (4-5 bp). With the exception of a few mutated copies, the newly identified elements have the same canonical dinucleotides terminating the LTRs as previously characterized HERVs (TG/CA). Since LTR_STRUC can only identify elements having two LTRs, we conducted BLAST searches using identified full-length elements as query sequences to identify solo LTRs and other fragmented elements. Consistent with what has been reported previously for other HERV families (Stoye 2001), we have found that each of the newly identified families are represented by significantly more solo LTRs and fragmented sequences than full-length elements.(Table 3.1).

Because HERV LTRs are synthesized from the same RNA template during reverse transcription, they are identical in sequence at the time of integration (Boeke and Stoye 1997). Using the primate pseudogene nucleotide substitution rate of 0.16% divergence / million years (Kapitonov and Jurka 1996; Costas and Naveira 2000; Jordan and McDonald 2002), the relative integration time or age of any full-length HERV can be estimated from the level of sequence divergence existing between the element's 5' and 3' LTRs. Using this method, the estimated age of the new families of HERVs described here

range from 18.0 to 49.5 MY indicating that members of these families have not been transpositionally active in the primate lineage since well before chimpanzees and humans diverged from a common ancestor (6 MYA) (Mikkelsen et al. 2005). Although caution must be taken when using LTR divergence to estimate the age of individual elements because of confounding processes such as recombination and conversion, [e.g., (Johnson and Coffin 1999; Hughes and Coffin 2005)], the method is able to provide useful age estimates, at least to a first approximation [e.g., (Bowen and McDonald 2001)]. Our estimated age of the newly identified human elements fall within the median range of previously described families of HERVs (Tristem 2000).

Table 3.1: Representative elements of human endogenous retroviral families characterized in this study (ND: Not Determined)

Family Name	Location on Chromosome [Chromosome No : Position (hg17)]	5' and 3' LTR identity	Length of 5' / 3' LTRs	Target site repeats	Dinucleotides	Element Length	tRNA primer	Age of element* (MY)	Copy Number#
HERV 1	10 : 42460513-42470423 (q11.21)	91.0	518/523	CCAC/CCAC	TG/CA	9911	Pro	30.0	~21
HERV 2	1 : 75864691..75868511 (p31.1)	76.0	328/329	TTTT/TTTT	TG/AA	3821	ND**	49.5	~8
HERV 4	4 : 75664267..75671483 (q13.3)	84	425/430	ACAG/ATAG	TG/CA	7217	Glu	56.0	~170
HERV 5	3 : 14055590..14063894 (p25.1)	90.7	359/361	TCAT/TCAT	TG/CA	8305	Gln	31.0	~27
HERV 6	15 : 82602937..82611762 (q25.3)	87.5	434/432	ND**	AG/CT	8826	ND**	42.7	~40
HERV 7	3 : 87738962..87748142 (p11.2)	88.7	659/682	ND**	TG/CA	9181	ND**	38.0	~36
HERV 10	6 : 120127640..120134837 (q22.31)	90.1	497/506	CAGT/CAGT	TG/CA	7198	ND**	33.0	~65
HERV 11	1 : 31599889..31609246 (p35.2)	94.0	622/622	GCAAA/GCAAA	TG/CA	9358	ND**	19.3	~67
HERV 12	3 : 168466792..168475512 (q26.1)	92.1	508/508	AGTT/AGTT	TG/CA	8721	ND**	25.9	~33

* Using the primate pseudogene nucleotide substitution rate of 0.16% divergence / million years (Kapitonov and Jurka 1996; Costas and Naveira 2000; Jordan and McDonald 2002), the relative integration time or age of full-length HERV was estimated from the level of sequence divergence existing between the element's 5' and 3' LTRs. The Jukes-Cantor model was used to correct for the presence of multiple mutations at the same site, back mutations and convergent substitutions.

****** Because of the accumulation of substitutions, it was not possible to accurately determine target site repeats and tRNA binding sites. # includes full length elements, fragmented copies and solo LTRs.

Table 3.2: Characteristics of human endogenous retroviral families identified in this study

Family Name	Length of Full elements	Length of LTRs	5' and 3' LTR identity	Ages of the elements (MY)
HERV 1	9671 to 10665	380 to 526	91.0 to 96.19	12.2 to 29.78
HERV 2	3821	328	76	49.52
HERV 4	7217 to 9766	249 to 489	82.5 to 92.62	24.25 to 62.24
HERV 5	8059 to 10724	359 to 637	87.9 to 92.14	25.92 to 41.22
HERV 6	8820 to 8825	430 to 434	87.0 to 88.0	42.0 to 43.0
HERV 7	9181 to 9668	563 to 682	88.29 to 88.7	38.01 to 39.75
HERV 10	7198 to 9400	495 to 509	90.1 to 94.4	18.0 to 33.0
HERV 11	9358 to 9916	552 to 622	90.0 to 94.0	19.38 to 30.16
HERV 12	8721 to 8794	484 to 508	88.42 to 92.4	24.97 to 25.99

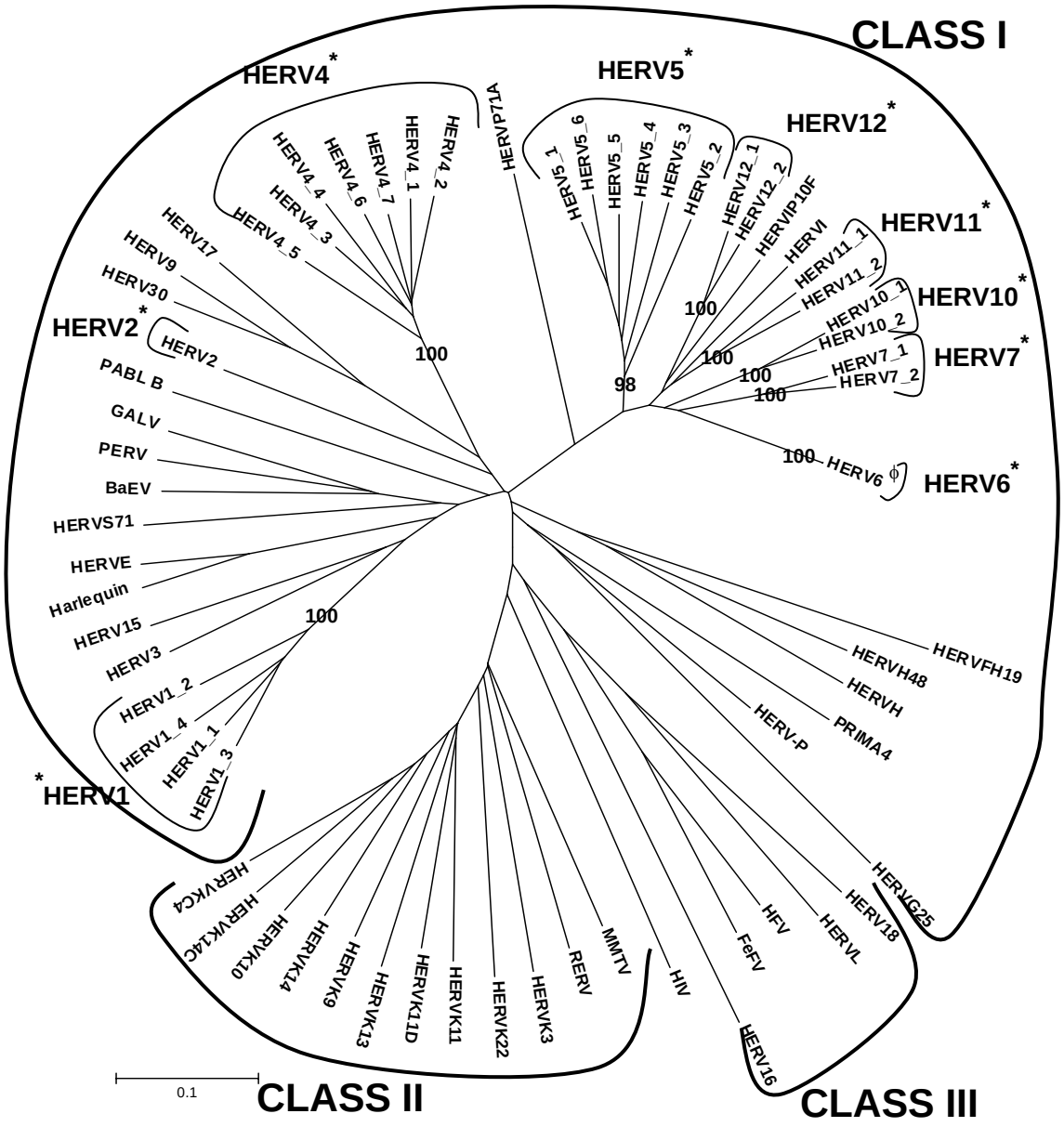

Figure 3: Unrooted RT based neighbour joining tree of human LTR retrotransposon families

The Phylogenetic tree is built from the DNA sequence [using MEGA3 software (Kumar et al. 2004)]of the reverse transcriptase taken from all the members of the newly identified HERV families together with a representative member of previously identified families. (* Indicates human LTR retrotransposon families characterized in this study, ϕ This element is present in the duplicated region of the genome on chromosome 15. As a result, six elements of this family of greater than 97 % identity are present in the human genome.). Elements are grouped into families based on bootstrap values (shown for families newly identified in this report). Previously identified families of retrotransposons and retroviruses are included for comparison (GALV: Gibbon ape leukemia virus [Accesion No: M26927], PERV : Porcine endogenous retrovirus [Accesion No : AF038601], BaEV : Baboon endogenous virus [Accesion No :

X05470], HFV : Human Foamy virus [Accesion No : Y07725], FeFV : Feline Foamy virus [Accesion No : AJ223851], HIV : Human immunodeficiency virus [Accesion No : K03454], MMTV : Mouse mammary tumor virus [Accesion No : NC_001503], RERV : Rabbit endogenous retrovirus [Accesion No : AF480925]).

ACKNOWLEDGEMENTS

This research was supported by a grant from the Georgia Institute of Technology

Research Foundation.

CHAPTER 4

DIFFERENTIAL GENE EXPRESSION PATTERNS BETWEEN HUMANS AND CHIMPANZEES IS ASSOCIATED WITH RETROTRANSPOSON INDEL VARIATION

ABSTRACT

It has been proposed that the genetic basis of the evolutionary divergence between humans and chimpanzees is the result of regulatory difference associated with the substantial insertion and deletion (INDEL) variation existing between the two species. To test this hypothesis we have categorized human-chimpanzee INDEL variation mapping in or near genes and determined if this variation is significantly correlated with species differences in gene expression. Our results indicate that the majority of this INDEL variation between humans and chimpanzees is associated with retrotransposon sequences and that this variation is significantly correlated with differences in gene expression most notably in the brain and testes. Our findings are consistent with the hypothesis that retrotransposons have played a significant role in human/chimpanzee evolution.

INTRODUCTION

Although humans and chimpanzees have accumulated significant differences in a number of phenotypic traits since diverging from a common ancestor about six million years ago, their genomes are >98.5% identical at protein coding loci (Mikkelsen et al. 2005). Since

this modest degree of nucleotide divergence does not seem sufficient to explain the extensive phenotypic differences that exist between the two species, it has been hypothesized that the genetic basis of the differences lies at the level of gene regulation (King and Wilson 1975) and possibly associated with the extensive INDEL (insertion/deletion) variation existing between the two species [e.g.,(Britten 2002)]. Recent studies have shown that retrotransposons may contribute significantly to the INDEL variation between humans and chimpanzees (Polavarapu et al. 2006b). Since retrotransposon sequences located in or near genes are known to have the capacity to significantly alter patterns of gene expression [e.g.,(Hasler and Strub 2006; Maksakova et al. 2006)], it has been postulated that these elements may be important factors in regulatory evolution [e.g., (McDonald 1993; Britten 1997; van de Lagemaat et al. 2003). In this paper, we present a detailed characterization of the INDEL variation (80-12,000 bp in length) associated with human and chimpanzee genes and correlate this variation with differences in gene expression in a variety of organs and tissues. The results are consistent with the hypothesis that retrotransposons have been significant players in human/chimpanzee evolution.

RESULTS AND DISCUSSION

We use the terms “human gaps” (HG) to refer to sequences present in chimpanzees but absent in humans and “chimpanzee gaps” (CG) to sequences present in humans but absent in chimpanzees (Polavarapu et al. 2006b). HGs and CGs together constitute the INDEL variation between humans and chimpanzees. Using the database available at the

UCSC Genome Bioinformatics web site (http://genome.ucsc.edu), we identified a total of 18,395 HGs and 27,728 CGs of which 10,220 (56%) and 16,541 (60%) contain interspersed repeats respectively. Nearly all of the HGs and CGs that contain interspersed repeats (~99%) are homologous to retrotransposon sequences (Table 4.1).

Table 4.1: Categories of Human and Chimpanzee gap sequences

Category	**Human Gaps**	**Chimpanzee Gaps**
Total Gaps	18395	27728
Gaps containing interspersed repeats (IRs)	10220 (55.55%)	16541 (59.65%)
IR gaps containing retrotransposons	10080 (98.63%)	16400 (99.14%)
SINE	4582 (45.5%)	9651 (59.0%)
LINE	2586 (25.5%)	2841 (17.0%)
ER	903 (9.0%)	850 (5.0%)
SVA	402 (4.0%)	926 (6.0%)
ME	1607 (16.0%)	2132 (13.0%)
DNA elements	140	141

The retrotransposon sequences associated with HGs and CGs can be grouped into five different classes: 1) SINEs (Short Interspersed Nuclear Elements), 2) LINEs (Long Interspersed Nuclear Elements) , 3) ERs (Endogenous Retroviruses), 4) SVAs (biologically active composite elements consisting of fragments of SINE R, VNTRs-variable number of tandem repeats, and Alu elements) and 5) MEs (Mosaic Elements-a term we will use to refer to those transpositionally inactive sequences comprised of a mosaic of more than one class of the above retrotransposon homologous sequences). Of the retrotransposon sequences associated with HGs, 45.5% are homologous to SINEs, 25.5% to LINEs, 9% to ERs, 4% to SVAs and 16% to MEs (Table 4.1). Of the retrotransposon sequences associated with CGs, 59% are homologous to SINEs, 17% to LINEs, 5% to ERs, 6% to SVA and 13% to MEs (Table 4.1). These values are

proportional to the relative frequency of the various classes of retrotransposons in the human and chimpanzee genomes (Lander et al. 2001; Mikkelsen et al. 2005).

The presence of a retrotransposon sequence in humans (or *vice versa* in chimpanzees) that is missing at an orthologous genomic position in chimpanzees (humans) can either be due to a novel insertion in one species or a deletion in the other. In most instances, it is possible to distinguish between these two events because a deleted retrotransposon sequence typically leaves a footprint of the original insertion event (e.g., target site duplications, solo LTRs in the case of ERs, etc [e.g., (Polavarapu et al. 2006b) and Material and Methods]). For each of the retrotransposon gaps existing between humans and chimpanzees, we have distinguished those due to insertions from those due to deletions. We found that the majority of the retrotransposon-associated INDEL variation between humans and chimpanzees (> 2-fold) is due to deletions [Sixty-five percent of retrotransposon associated sequence gaps in humans are the result of deletions while 35% are the consequence of insertions in chimpanzees. Similarly, 75% of the retrotransposon sequence gaps in chimpanzees are due to deletions and 25% to insertions in humans (Table 4.2)].

Consistent with the relative transpositional activity of retrotransposon families in humans and chimpanzees (Lander et al. 2001; Mikkelsen et al. 2005), we found that the majority of the insertions involve SINEs and LINEs (Table 4.2). The frequency of ER insertions is > 2 fold higher in chimpanzees than in humans predominately due to the expansion of two chimpanzee-specific ER families (CERV 1/PTERV1 and CERV 2) 3-5 million years

ago (Yohn et al. 2005; Maksakova et al. 2006; Polavarapu et al. 2006b). In contrast, we found that the frequency of SVA insertions in humans is > 2-fold higher than in chimpanzees (Table 4.2). The overall frequency of retrotransposon sequence deletions is nearly 2-fold higher in chimpanzees than in humans. The frequency of chimpanzee SINE and SVA deletions is nearly 3-fold higher than humans while the frequency of ER deletions in the two species is nearly identical (Table 4.2).

Table 4.2: Retrotransposon insertions and deletions in humans and chimpanzees

	Human retrotransposon homologous gaps		**Chimpanzee retrotransposon homologous gaps**	
	10080		16400	
	Chimpanzee Insertions	Human Deletions	Human Insertions	Chimpanzee Deletions
Total	3451 (34.2%)	6629 (65.8%)	3918 (24.0%)	12482 (76.0%)
SINE	2366	2216	2929	6722
LINE	807	1779	658	2183
ER	151	752	70	780
SVA	127	275	261	665
ME	0	1607	0	2132

We found that 38% of human and chimpanzee genes (RefSeq) are associated with gaps (HG and/or CG) within or in proximity (1500bp upstream or downstream) to genes. Of these, 75% are associated with retrotransposon associated gaps. Eighty-two percent of genes associated with retrotransposon associated gaps are due to deletions while 37% are due to insertions (19% of the genes were associated with both deletions and insertions in either of the species). Thus, the vast majority of the human-chimpanzee retrotransposon associated INDEL variation located in or in proximity to genes is due to a lineage-specific loss of sequences that were present in their common ancestor > 6 MYA.

To explore the relationship between human-chimpanzee retrotransposon INDEL variation with differences in gene expression, we reanalyzed a previously published human-chimpanzee expression dataset consisting of expression arrays from five different tissues-brain, heart, liver, kidney and testis (Khaitovich et al. 2005). In the previous study, the data were used to correlate sequence differences with expression differences and, as a consequence, many probe sets were discarded for which quality sequence was not available for both species. In our re-analysis, these probe sets were included since detailed sequence information is not needed to correlate INDEL variation with differences in gene expression. Our re-analysis indicates that the most dramatic difference in gene expression between humans and chimpanzees is in testis (60% of genes display a significant difference in expression) followed by brain (35%), heart (35%), kidney (32%) and liver (25%) (Table 4.3). Significant gene expression differences between humans and chimpanzees in brain and testes have been previously noted by several authors [e.g., (Enard et al. 2002; Caceres et al. 2003; Khaitovich et al. 2005)].

Table 4.3: Number of genes differentially expressed between humans and chimpanzees in different tissues.

	Brain	**Heart**	**Liver**	**Kidney**	**Testis**
Expressed genes	10231	9580	9451	10546	12509
Genes differentially expressed	3519 (34.4%)	3361 (35.1%)	2385 (25.2%)	3458 (32.8%)	7783 (62.2%)
Genes not Differentially expressed	6712 (65.6%)	6219 (64.9%)	7066 (74.8%)	7088 (67.2%)	4726 (37.8%)

Of the ~ 25,000 genes examined in the microarray analysis, 64% displayed significant differences in gene expression in at least one of the five tissues examined. Of the 7026

genes associated with human-chimpanzee INDEL variation, 76% displayed highly significant differences in gene expression between the two species compared to genes not associated with INDEL variation (p = 2.2e-16). The correlations were highly significant for differences in expression in brain (p = 5.26e-05), liver (p=1.97e-09) and testes (p=4.28e-05) (Table 4.4). We found that 77% of the 5193 genes associated with retrotransposon associated INDEL variation are significantly differentially expressed in brain (p = 0.00013), heart (p = 7.79e-05), kidney (p = 2.69-07) and testis (p = 0.0027) while genes not associated with retrotransposon INDEL variation are significantly differentially expressed only in kidney (p =0.0068) (Table 4.4).

Table 4.4: Correlation (p-values) between INDEL variation and differences in human-chimpanzee gene expression patterns

	Brain	**Heart**	**Liver**	**Kidney**	**Testis**
Indel Variation	5.26e-05	0.0016	0.467	1.97e-09	4.28e-05
Retrotransposon Indel Variation	0.00013	7.79e-05	0.33	2.69e-07	0.0027
Non retrotransposon Indel Variation	0.20	0.81	0.312	0.0068	0.029

In order to determine if the location of an INDEL in or near genes may have a significant effect on expression differences, we grouped genes with respect to the position of the associated INDELS into four categories: 1) Exon, 2) Upstream (within 1500bp upstream of transcription start site), 3) Downstream (within1500bp downstream of transcription termination site), and 4) Intron. As might be expected, relatively little human/chimpanzee INDEL variation maps to exons. Many (but not all) of the INDELS falling in this category involve relatively few bases or sequences located in non-encoding regions (Table 4.5). We found that genes associated with all types of INDEL variation mapping

to exons display differential gene expression only in kidney (p = 0.0007) while genes associated specifically with retrotransposon INDEL variation mapping to exons display differential gene expression in heart (p = 0.0064), kidney (p=0.0011) and testis (p=0.014) (Table 4.6).

Table 4.5: Indel variation associated with exons

	HGs associated with exons	**CGs associated with exons**
Total	116	130
Coding	45 (6 involve only 1 bp)	75 (3 involve 1 bp)
Non coding (5'&3' UTRs)	71	55

Table 4.6: Correlation (p-values) between INDEL variation located in exon and intron of genes and human-chimpanzee differential gene expression

TISSUE								
	EXON				INTRON			
	ALL GAPS		RETRO GAPS		ALL GAPS		RETRO GAPS	
	Exp	Df exp	Exp	Df exp	Exp	Df exp	Exp	Df exp
BRAIN	96	31	31	13	3703	1362***	2942	1088***
TESTES	127	82	36	30**	4225	2744***	3336	2157***
HEART	96	41	34	20**	3383	1255**	2676	1023***
LIVER	96	25	31	11	3247	838	2534	663
KIDNEY	112	54***	35	21***	3693	1338***	2903	1064***

*p<0.05	**p<0.01	***p<0.001

Genes associated with human-chimpanzee INDEL variation located within 1.5 kb upstream of the transcriptional start site were not significantly differentially expressed in any tissue (Table 4.7). This lack of significance is, in large measure, due to the fact that there are relatively few gaps located in this 1.5 kb region. Since biologically important

control sequences are known to map within 1.5 kb of eukaryotic genes, a possible explanation of our results is that most functionally significant INDELs arising in this region were removed by natural selection and most, if not all, of the remaining INDEL variation in this region is of little functional consequence. We observed that INDEL variation mapping to downstream flanking regions was marginally associated with differences in gene expression in heart (p =0.05), liver (p = 0.03) and kidney (0.03) (Table 4.7).

Table 4.7: Correlation (p-values) between INDEL variation located in upstream and downstream of genes and human-chimpanzee differential gene expression

TISSUE								
	UPSTREAM				DOWNSTREAM			
	ALL GAPS		RETRO GAPS		ALL GAPS		RETRO GAPS	
	Exp	Df exp	Exp	Df exp	Exp	Df exp	Exp	Df exp
BRAIN	178	62	99	38	212	84	106	38
TESTES	217	140	121	75	256	148	130	71
HEART	169	56	95	25	203	85*	101	41
LIVER	175	47	103	20	197	63*	94	30
KIDNEY	190	71	109	39	204	82*	100	36

*p<0.05	**p<0.01	***p<0.001

Interestingly, the class of human-chimpanzee INDEL variation found to be most significantly associated with differences in gene expression maps to introns [all INDEL variation: brain (p=0.00014), heart (p= 0.0024), kidney (p=3.71e-08) and testis (p=7.67e-06); retrotransposon INDEL variation: brain (p=0.0005), heart (p=6.5e-05), kidney (p=2.8e-07) and testis (p=0.0007)] (Table 4.6). Several recent molecular evolutionary studies have documented the incorporation of intronic retrotransposon sequences into

exons by virtue of the presence of cryptic splice-donor sequences in the retrotransposon [e.g.,(Kreahling and Graveley 2004; DeBarry et al. 2006)]. We identified 3610 genes with retrotransposon sequence gaps in chimpanzee introns (i.e., retrotransposon sequences that are present within the introns of human genes but missing within the introns of the orthologous chimpanzee genes). We looked for the incorporation of these sequences into human mRNAs by searching the repository of 222,757 validated human mRNAs available at UCSC genome browser (http://genome.ucsc.edu) and found that 232 or 6.4% of human genes containing retrotransposon sequences within their introns (242 distinct sequences) have incorporated these sequences into their mRNAs (Table 4.8).

Table 4.8: Human mRNA sequences associated with chimpanzee gaps

	Human genes (CGs)	**Human genes (distinct mRNA sequences)**	
CGs	6551 (13,302)	407 (463)	
Retrotransposon associated CGs	3610 (8105)	232 (242)	
		Human Insertions (43)	Chimpanzee Deletions (199)
SINEs		35	92
LINEs		4	15
ERs		2	13
SVAs		2	19
MEs		0	60

As is the case for all retrotransposon associated INDEL variation between humans and chimpanzees, the majority (199/242 or 82%) of incorporated intronic sequences present in humans but absent in chimpanzees is due to deletions. The remaining 18% (43/242) is the result of insertions (predominately SINES) into human introns (Table 4.8). We conclude that although some of the gene expression variation between humans and chimpanzees associated with intronic gaps may well be due to the incorporation of

retrotransposon sequences into mRNAs (affecting RNA stability, etc.), the majority of these intronic sequences may be exerting enhancer or other cis-regulatory controls on transcription. Several examples of such retrotransposon-mediated cis-regulatory effects have been documented in mice, rats and humans [e.g., (Rothenburg et al. 2002; Yamada et al. 2006; Illarionova et al. 2007)] .

We analyzed the expression levels of genes associated with the various classes of retrotransposon sequences (SINEs, LINEs, ERs, SVAs and MEs) to determine if any particular class was preferentially associated with genes displaying significant differences in expression between humans and chimpanzees. We found that genes associated with SINE-generated INDEL variation are differentially expressed between humans and chimpanzees in brain (p=0.00064), heart (p=0.0028), kidney (p=1.66e-05) and testis (p=0.010) while genes associated with other classes of retrotransposons are significantly differentially expressed in relatively few tissues (e.g., genes associated with LINEs only in heart, SVAs only in testis, and mosaic elements only in brain and kidney) (Table 4.9). SINES are the most abundant and transpositionally active class of retrotransposons in humans and chimpanzees and have frequently been associated with functionally important insertional mutations in humans [e.g.,(Deininger and Batzer 1999; Hasler and Strub 2006)].

Table 4.9: Correlation (p-values) between retrotransposon indel variation and differences in human-chimpanzee gene expression

	Brain	**Heart**	**Liver**	**Kidney**	**Testis**
SINEs	0.00064	0.0028	0.32	1.66e-05	0.010
LINEs	0.678	0.00038	0.577	0.689	0.824
ERs	0.59	0.108	0.714	0.68	0.22
SVAs	0.073	0.99	0.52	0.06	0.017
Mosaic	0.020	0.288	0.47	0.013	0.47

Among the phenotypic differences that distinguish humans from chimpanzees, differences in fertility and cognitive abilities have been singled out as being of particular evolutionary importance [e.g., (Eberhard 1985; Tomasello and Call 1997; Wyckoff et al. 2000; Gu and Gu 2003)]. Interestingly, we found that a number of the genes associated with retrotransposon INDEL variation and displaying significant differences in expression in brain and testes have been previously associated with critical functions (Table B.1, B.2). For example, many of the genes associated with retrotransposon sequence gaps that display significant differences in expression between human and chimpanzee brains are known to be involved in nervous system development (e.g., *CHD5, DSCAM, NTRK2*), memory and learning *(*e.g., *GRIN2A, IL1RAPL1)*. Other differentially expressed genes associated with retrotransposon sequence gaps have been previously implicated in brain related diseases in humans [e.g., Parkinson and Alzheimer's disease *(SNCA),* Hirschsprung Disease Syndrome *(ZFHX1B),* Bardet-Biedl syndrome *(BBS2),* Niemann-Pick Disease *(NPC1)* and Riley-Day syndrome *(FD, IKBKAP)*] (Table B.1). Likewise, we found that a large number of retrotransposon associated genes that display a significant difference in expression between human and chimpanzee testes have been implicated in spermatogenesis (*PBK, TESK2, DDX4, SUV39H2, CLGN, ADAM18, MAP7*), male infertility (*USP9Y CLGN, GPR64*), the resetting of methylation marks during male germ cell differentiation (*CTCFL*) and spermatogenesis (*TSKS, ADAM18*) (Table B.2).

Over the ~ 6 million years that the human and chimpanzee lineages have been diverged from a common ancestor, the two species evolved a variety of distinctive morphological, behavioral, cognitive and other phenotypic traits (Varki and Altheide 2005). To explore the genetic basis of the phenotypic differences that distinguish humans from chimpanzees, a number of comparative genomic studies have been conducted in recent years [e.g., (Li et al. 2001; Mikkelsen et al. 2005). Perhaps the most surprising finding coming out of these studies is the paucity of nucleotide variation existing between these two species supporting earlier contentions that the basis of the phenotypic differences lies in the realm of gene regulation (King and Wilson 1975). Direct evidence in support of the regulatory hypothesis has recently been provided by a number of comparative microarray studies showing that significant differences in gene expression patterns exist between humans and chimpanzees especially in organs (e.g., brain and testes) and functions (e.g., cognitive ability and fertility) directly related to some of the major phenotypic traits distinguishing the two species (Varki and Altheide 2005). Questions remain, however, concerning the genetic basis of the differences in gene regulation that separates humans from chimpanzees. One recently offered hypothesis is that the substantial INDEL variation that exists between humans and chimpanzees may contribute significantly to the regulatory differences between the species (Britten 2002; Polavarapu et al. 2006b). To test this hypothesis, we have categorized the INDEL variation existing between humans and chimpanzees that is located in or near genes and determined if this variation is significantly correlated with species differences in gene expression. Our results indicate that such a correlation does indeed exist and that it is primarily attributable to retrotransposon associated INDEL variation. Interestingly, the majority of this variation

is attributable to lineage specific deletions. Evidence for "bursts" of retrotransposon activity within the primate lineage pre-dating the humans-chimpanzee divergence have been presented by several authors [e.g., (Sverdlov 2000; Boissinot and Furano 2005)]. Our results suggest a model whereby such periods of high genomic instability may have provided some of the "raw material" used by natural selection to shape human-chimpanzee regulatory differences by means of lineage-specific deletion events.

Once considered parasitic or "junk" DNA (Doolittle and Sapienza 1980) of little or no functional significance, transposable elements have, in recent years, been recognized as significant contributors to regulatory variation both within and between species [e.g.,(Medstrand et al. 2005)]. Our results are generally consistent with these findings and, in particular, indicate that retrotransposon mediated regulatory variation may have been a significant factor in human/chimpanzee evolution.

METHODS

Initial dataset

We utilized HG and CG datasets available at the UCSC Genome Bioinformatics web site (http://genome.ucsc.edu) that were generated by aligning the chimpanzee genome with human genome build HG16 (Karolchik et al. 2003; Karolchik et al. 2004). These datasets include gaps of sizes ranging from 80 bp to 12 kb. The Repeat Masker program was used (AF Smit and P Green, unpublished data) to identify all interspersed repeats, i.e. all transposable elements present in the datasets, and retrotransposon homologous sequences were subsequently separated. The CG dataset coordinates were later updated to HG18

version of the human genome using genome browser utilities, Batch Coordinate Conversion (liftOver) tool (http://genome.ucsc.edu/util.html). A few sequences (76) not represented in the newer version of human genome (HG18) were removed during this process.

Separation of retrotransposon homologous gap sequences to insertions and deletions

The HG and CG sequences homologous to interspersed repeats were divided into two types: 'Single type' gap sequences are defined as those composed of only sequences homologous to one class of interspersed repeats i.e either SINE, LINE, ER, SVA or DNA elements; and 'Mosaic type' gap sequences are defined as those composed of more than one category of interspersed repeats (for example, ER inserted within a LINE element). Single type gap sequences were further divided into two categories: category 1 comprises gap sequences composed entirely of a retrotransposon sequence; and category 2 comprises gap sequences composed of retrotransposon and non-interspersed repeat sequences. The above categorizations were useful for distinguishing gaps due to deletions in one species from gaps due to insertions in the other species. Instances of mosaic type and single type category 2 gaps were deletions in the species with gaps while gaps belonging to single type category 1 were either deletions in the species with gaps or insertions in the other species.

Because retrotransposon sequences do not excise precisely (Boeke and Stoye 1997) it is possible to distinguish between these two possibilities. If a region in humans orthologous to the position of retrotransposon sequence in chimpanzees contains a remnant of

retrotransposon sequence (e.g., fragmented element, solo LTR etc), we score the gap as a deletion in humans. If the orthologous region contains no remnant of retrotransposon sequence, we score the gap as an insertion in chimpanzees. The same rules apply for analogous dataset of retrotransposon sequences present in humans but absent in chimpanzees. Although precise deletions of Alu sequences were reported (van de Lagemaat et al. 2005), and our criteria of separating insertions from deletions would incorrectly identify these cases as insertions, such events were very rare (0.5%-1%) (van de Lagemaat et al. 2005) and since our dataset is huge they do not influence the results we report here.

Identification of human and chimpanzee genes associated with INDEL variation

The coordinates for human and chimpanzee RefSeq genes (Transcription start, transcription end, coding sequence (CDS) start, CDS end, exon start, exon end etc) were downloaded from UCSC Genome Bioinformatics web site (http://genome.ucsc.edu)(Karolchik et al. 2004). It is known from previous molecular studies that most regulatory regions of human genes are located within the gene or 1500bp upstream or downstream of the gene. A gap is considered as associated with the gene if it is present in this region. In-house perl scripts were written to match coordinates of HG and CG sequences with RefSeq gene coordinates. We found a total of 4017 human Refseq genes and 5162 chimpanzee Refseq genes to be associated with HG and CG sequences. Of these, 2560 and 3836 genes were associated human retrotransposon homologous gaps and chimpanzee retrotransposon homologous gaps respectively.

Gene Expression data analysis

The human-chimpanzee gene expression data from five different tissues (i.e. brain, heart, liver, kidney and testis) in 6 humans and 5 chimpanzees were obtained from a previous study (Khaitovich et al. 2005). The samples were studied using Affymetrix (Santa Clara, California, United States) HGU133plus2 arrays. The expression data were reanalyzed using the following procedure. The data were normalized using MAS normalization method and genes with detection p-values of less than 0.065 were considered as detected in a given tissue. The genes with significant sequence differences in Affy probes between humans and chimpanzees and with inconsistent hybridization patterns within samples in a species were removed. Analysis of Variance (ANOVA) was used to obtain genes significantly differentially expressed (genes with p-values lesser than 0.01) between the two species.

Categories of genes associated with INDEL variation between humans and chimpanzees

Genes associated with HG and CGs were analyzed in two different ways: 1) Based on type of gap sequences, whether it is homologous to a retrotransposon sequence or not. For this analysis, we divided INDEL variation dataset into two different categories: a) retrotransposon INDEL variation and c) non retrotransposon INDEL variation, and 2) Based on location of the INDEL variation i.e. whether INDEL variation is present in intron, exon, upstream (within 1500bp upstream of transcription start site) or downstream

(within1500bp downstream of transcription termination site) of a gene. Some genes were associated with IV in two or more regions, i.e gap starting upstream of gene and ending in first intron. Such genes were included in more than one category depending on the regions covered by gap sequences. The genes associated with retrotransposon INDEL variation were further divided based on retrotransposon class, whether the sequence is homologous to SINE, LINE, ER, SVA or MEs. As with the previous analysis, some genes were associated with many gap sequences each of which is homologous to a different class of retrotransposon sequences. Such genes were included in more than one category depending on the number of retrotransposon classes contained in the gap sequences.

Correlating INDEL variation with differential expression

The genes in each of the above defined categories were checked for their expression levels between humans and chimpanzees in each of the five tissues. We used the same criteria described above in considering a gene as detected or differentially expressed between humans and chimpanzees. All genes that were detected but not differentially expressed were considered as not differentially expressed between humans and chimpanzees. We used statistical package in R for obtaining statistical significance of differential expression of genes associated with different categories of INDEL variation. The expression levels of genes without INDEL variation in a given categories were used as background sets against which expression levels of genes associated with INDEL variation were compared. We used proportions test for obtaining statistical significance

of above comparisons and cases with p-values of less than 0.05 were considered as statistically significant.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Georgia Institute of Technology Research Foundation.

CHAPTER 5

CONCLUSION

In conclusion, this research outlays systematically for the first time the genomic differences existing between humans and chimpanzees with respect to retrotransposon sequences and contributes significantly to the understanding of the possible role these elements played in human and chimpanzee evolution, in particular, evolution of gene regulation between humans and chimpanzees. Once considered parasitic sequences of little or no adaptive significance [(Doolittle and Sapienza 1980; Orgel and Crick 1980)], transposable elements are today generally recognized as significant contributors to human regulatory [e.g., (Jordan et al. 2003)] and structural [e.g., (Nekrutenko and Li 2001)] gene evolution. . The sequencing of the chimpanzee genome (Mikkelsen et al. 2005) and the availability of expression data from five different tissues (*viz* brain, heart, liver, kidney, testis) in humans and chimpanzees (Khaitovich et al. 2005) has provided an unprecedented opportunity to not only compare the full complement of retrotransposons in two closely related primate species but to gain insight into the role these elements may have played in human/chimpanzee evolution.

A first systematic search for one particular class of retrotransposons—the endogenous retroviruses (ERVs), was carried out in the chimpanzee genome. ERVs are genomic elements that are very similar to the more familiar infectious retroviruses, such as HIV, but they are unable to move from cell-to-cell. These elements are thought to be remnants of ancient germline infections, and are capable of transposing within the genome by

encoding regulatory features including transcriptional promotion and termination signals. The transposition events lead to the accumulation of ERV sequences in the genome and, along with the elimination of element sequences, leading to pronounced regulatory differences between evolutionary lineages.

Forty two families of ERVs were identified in the chimpanzee genome including the discovery of 9 previously unknown families in humans. The members of these families range in age from about 0.8 to 145 MY. The vast majority of these families were found to have orthologous, i.e. elements in corresponding genomic positions, in the human genome except for two (CERV 1/PTERV1 and CERV 2) families. The presence of orthologous families indicates that these elements were around prior to the diversification of the two species. Nevertheless, nine families of chimpanzee ERVs have been transpositionally active since the human-chimpanzee divergence, while only two families have been active along the human lineage. Thus, while some families of endogenous retroviruses have not been transpositionally active within the primate lineage, others have and continued to be active since chimpanzees and humans diverged from a common ancestor. The two CERV families without orthologs in the human genome display a patchy distribution among primates and our data suggest that at least some members of both families have been transpositionally active in the chimpanzee lineage after the divergence of chimpanzees and humans from a common ancestor. Several lines of evidence indicate that CERV 1/PTERV1 and CERV 2 elements arose in the chimpanzee genome from exogenous infection and that humans developed resistance to such infections.

The nine previously unidentified families of Human Endogenous Retroviruses (HERVs) are characterized in detail. All are low abundance families being comprised of only 1-7 full-length members with low homology to previously identified HERVs. Each of the newly identified families is represented by significantly more solo LTRs and fragmented sequences than full-length elements. The estimated age of these families range from 18.0 to 49.5 MY indicating that members of these families have not been transpositionally active in the primate lineage since well before humans and chimpanzees diverged from a common ancestor (6 MYA).

It has previously been reported that the biggest differences between the human and chimpanzee genomes result from insertion and deletion (INDEL) events (3.5%) rather than point substitutions (1.5%)(Britten 2002). One recently offered hypothesis is that this INDEL variation may contribute significantly to the regulatory variation between the species (Britten 2002; Frazer et al. 2003; Polavarapu et al. 2006b) which has been hypothesized to be the basis of the distinctive morphological, behavioral, cognitive and other phenotypic traits (Varki and Altheide 2005) existing between the two species. A preliminary survey of endogenous retroviral positional variation between humans and chimpanzees determined that ~7% of all human-chimpanzee INDEL variation is associated with endogenous retroviral sequences.

A detailed characterization of the INDEL variation associated with human and chimpanzee genes was conducted and was correlated with differences in gene expression

patterns in a variety of organs and tissues in an effort to identify its significance. The results indicated that the majority (60%) of INDEL variation between humans and chimpanzees is associated with retrotransposon sequences and that this variation is significantly correlated with differences in gene expression most notably in brain and testes. These results are generally consistent with the findings that retrotransposons are significant contributors to regulatory variation both within and between species [e.g., (Britten 1996; Brosius 1999a)]. In particular, these results indicate that retrotransposon mediated regulatory variation may have been a significant factor in human/chimpanzee evolution.

The ultimate regulatory contribution of retrotransposons on a particular gene is difficult to discern. Identifying and analyzing potential genes is a first step, and this dissertation has identified a number of targets, particularly in brain and testis, with possible retrotransposon mediated gene expression differences. Further molecular and population studies on identified targets will lead to a better understanding of retrotransposon contribution to human/chimpanzee regulatory gene evolution.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

CERV families with previously unrecognized human orthologues

Nine novel Class I CERV families with previously unidentified orthologues in the human genome are characterized in this paper for the first time (Table 2.1).

CERV 6: We identified ~21 full-length elements, fragmented copies and solo LTRs of CERV 6 in the chimpanzee genome. CERV 6 elements range in size from 9.6 – 10.1 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 4 – 5 bp TSD (Table 2.1). CERV 6 elements have a threonine tRNA primer binding site (Table 2.1). The LTRs of the CERV 6 family of elements range from 511 to 649 bp in length. Based on their LTR sequence identity (92.16 % to 96.53 %), we estimate that these CERV 6 elements inserted into the primate lineage between 10.87 –25.86 MYA, i.e., well before the divergence of chimpanzees and humans from a common ancestor.

CERV 9**:** CERV 9 elements are among the oldest LTR retrotransposons in the chimpanzee genome. Approximately 8 full-length elements, fragmented copies and solo LTRs of CERV 9 were found in the chimpanzee genome. CERV 9 elements range in size from 3.8 – 4.2 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 5 bp TSD (Table 2.1). CERV 9 elements have a histidine tRNA primer binding site (Table 2.1). The LTRs of the CERV 9 family of elements range from 195 to 318 bp in length. Based on their LTR sequence identity (68.63 % to 88.36 %), we estimate that

these CERV 9 elements inserted into the primate lineage between 36.38 –98.05 MYA.

CERV 14**:** Approximately 170 full-length elements, fragmented copies and solo LTRs of CERV 14 were found in the chimpanzee genome. CERV 14 elements range in size from 7.1 – 9.8 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 4 bp TSD (Table 2.1). CERV 14 elements have either a leucine or arginine tRNA primer binding site (Table 2.1). The LTRs of CERV 14 family of elements range from 305 to 459 bp in length. Based on their LTR sequence identity (80.5 % to 86.42 %), we estimate that these CERV 14 elements inserted into the primate lineage between 42.44 –60.93 MYA.

CERV 21**:** Approximately 27 full-length elements, fragmented copies and solo LTRs of CERV 21 were identified in the chimpanzee genome. CERV 21 elements range in size from 8.0 – 10.8 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 4 bp TSD (Table 2.1). CERV 21 elements have either a threonine or a proline tRNA primer binding site (Table 2.1). The LTRs of CERV 21 family of elements range from 364 to 637 bp in length. Based on their LTR sequence identity (87.8 % to 91.2 %), we estimate that these CERV 21 elements inserted into the primate lineage between 27.5 – 37.9 MYA.

CERV 22**:** Approximately 40 full-length elements, fragmented copies and solo LTRs of CERV 22 were found in the chimpanzee genome. CERV 22 elements range in size from 5.9 – 8.5 kb in length, are bordered by inverted terminal repeats (AG and CT). Because of the accumulation of substitutions at the borders of two elements in this family, it was not possible to accurately determine TSDs for this family (Table 2.1). CERV 22 elements have a threonine tRNA primer binding site (Table 2.1). The LTRs of CERV 22 elements range

from 256 to 431 bp in length. Based on their LTR sequence identity (86.6 % to 90.6 %), we estimate that these CERV 21 elements inserted into the primate lineage between 29.5 – 41.9 MYA.

CERV 23**:** We identified approximately 36 full-length elements, fragmented copies and solo LTRs of CERV 23 in the chimpanzee genome. CERV 23 elements range in size from 9.2 – 9.8 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 4 bp TSD (Table 2.1). CERV 23 elements have a proline tRNA primer binding site (Table 2.1). The LTRs of CERV 23 elements range from 575 to 681 bp in length. Based on their LTR sequence identity (84.7 % to 87.6 %), we estimate that these CERV 23 elements inserted into the primate lineage between 38.6 – 47.8 MYA.

CERV 24**:** Approximately 65 full-length elements, fragmented copies and solo LTRs of CERV 24 were found in the chimpanzee genome. CERV 24 elements range in size from 8.6 – 9.1 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 4 bp TSD (Table 2.1). CERV 24 elements have a proline tRNA primer binding site (Table 2.1). The LTRs of the CERV 24 family of elements range from 427 to 437 bp in length. Based on their LTR sequence identity (88.2 % to 89.1 %), we estimate that these CERV 24 elements inserted into the primate lineage between 34.0 – 37.0 MYA.

CERV 25**:** Approximately 67 full-length elements, fragmented copies and solo LTRs of CERV 25 were found in the chimpanzee genome. CERV 25 elements range in size from 9.5 – 9.9 kb in length, are bordered by inverted terminal repeats (TG and CA) and have 5 bp TSD (Table 2.1). CERV 25 elements have a proline tRNA primer binding site (Table

2.1). The LTRs of the CERV 25 family of elements range from 549 to 617 bp in length. Based on their LTR sequence identity (90.3 % to 94.3 %), we estimate that these CERV 25 elements inserted into the primate lineage between 17.91 - 30.34 MYA.

CERV 26**:** Approximately 33 full-length elements, fragmented copies and solo LTRs of CERV 26 were found in the chimpanzee genome. CERV 26 elements range in size from 9.3 – 10.9 kb in length, are bordered by inverted terminal repeats (TG and CA) and 4 bp TSD (Table 2.1). Because of sequencing ambiguities, it was not possible to determine the tRNA binding site for CERV 26 elements (Table 2.1). The LTRs of the CERV 26 family of elements range from 494 to 509 bp in length. Based on their LTR sequence identity (93.4 % to 93.9 %), we estimate that these CERV 26 elements inserted into the primate lineage between 19.1 – 20.7 MYA.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Table B.1: List of selected genes significantly differentially expressed in brain and associated with retrotransposon INDEL variation

Gene Symbol	p-value of differential expression	Human Gap Location	Chimp Gap location	Gene Function
CNTNAP2	4.90E-06	INTRON	INTRON	May play a role in the formation of functional distinct domains critical for saltatory conduction of nerve impulses in myelinated nerve fibers. Seems to demarcate the juxtaparanodal region of the axo-glial junction. Predominantly expressed in nervous system
CTNND2	7.58E-06	INTRON	INTRON	Functions as a transcriptional activator when bound to ZBTB33. May be involved in neuronal cell adhesion and tissue morphogenesis and integrity by regulating adhesion molecules. Predominantly expressed in brain.
SNCA	2.50E-05	DOWNSTREAM	INTRON	Alpha-synuclein is a member of the synuclein family, which also includes beta- and gamma-synuclein. Synucleins are abundantly expressed in the brain and alpha- and beta-synuclein inhibit phospholipase D2 selectively. SNCA may serve to integrate presynaptic signaling and membrane trafficking. Defects in SNCA have been implicated in the pathogenesis of Parkinson disease. SNCA peptides are a major component of amyloid plaques in the brains of patients with Alzheimer's disease.
CHD5	9.52E-05	INTRON	DOWNSTREAM	May play a role in the development of the nervous system and the pathogenesis of neural tumors.
DSCAM	0.0003442	INTRON	INTRON	Cell adhesion molecule that can mediate cation- independent homophilic binding activity. Could be involved in nervous system development. Primarily expressed in brain.

CDH8	0.000406	INTRON	INTRON	Cadherins are calcium dependent cell adhesion proteins. They preferentially interact with themselves in a homophilic manner in connecting cells; cadherins may thus contribute to the sorting of heterogeneous cell types. This particular cadherin is expressed in brain and is putatively involved in synaptic adhesion, axon outgrowth and guidance.
DLGAP1	0.0006014	INTRON	INTRON	Part of the postsynaptic scaffold in neuronal cells. Expressed in brain.
SNTG1	0.0008767	INTRON	INTRON	Adapter protein that binds to and probably organizes the subcellular localization of a variety of proteins. May link various receptors to the actin cytoskeleton and the dystrophin glycoprotein complex (By similarity). May participate in regulating the subcellular location of diacylglycerol kinase-zeta to ensure that diacylglycerol is rapidly inactivated following receptor activation. Brain specific. In CNS, it is expressed in the perikaryon and proximal portion of the neuronal processes. Strong expression in the hippocampus, neuron-rich dendate granule cells, and pyramidal cell layers. Highly expressed in neurons of the cerebral cortex. Also expressed in the cerebellar cortex, deep cerebellar nuclei, thalamus, and basal ganglia.
DOK6	0.0009678	INTRON	INTRON	Docking proteins interact with receptor tyrosine kinases and mediate particular biological responses. DOK6 promotes Ret-mediated neurite growth. May have a role in brain development and/or maintainance. Highly expressed in fetal and adult brain. Highly expressed in the cerebellum.
NCAM2	0.0010859	INTRON	INTRON	May play important roles in selective fasciculation and zone-to-zone projection of the primary olfactory axons. Expressed most strongly in adult and fetal brain
GRIN2A	0.0018922	INTRON	N/A	N-methyl-D-aspartate (NMDA) receptors are a class of ionotropic glutamate receptors. NMDA channel has been shown to be involved in long-term potentiation, an activity-dependent increase in the

				efficiency of synaptic transmission thought to underlie certain kinds of memory and learning. NMDA receptor channels are heteromers composed of the key receptor subunit NMDAR1 (GRIN1) and 1 or more of the 4 NMDAR2 subunits: NMDAR2A (GRIN2A), NMDAR2B (GRIN2B), NMDAR2C (GRIN2C), and NMDAR2D (GRIN2D
ZFHX1B	0.0020006	INTRON	N/A	Transcriptional inhibitor that binds to DNA sequence 5'- CACCT-3' in different promoters.Defects in ZFHX1B are a cause of Hirschsprung disease syndrome [MIM:235730]; also called Mowat-Wilson syndrome. Hirschsprung disease syndrome is an autosomal dominant complex developmental disorder. Individuals with functional null mutations present with mental retardation, delayed motor development, epilepsy, and a wide spectrum of clinically heterogeneous features suggestive of neurocristopathies at the cephalic, cardiac, and vagal levels
NRXN3	0.0028781	INTRON	INTRON	Neurexins are a family of proteins that function in the vertebrate nervous system as cell adhesion molecules and receptors. They are encoded by several unlinked genes of which two, NRXN1 and NRXN3, are among the largest known human genes. Three of the genes (NRXN1-3) utilize two alternate promoters and include numerous alternatively spliced exons to generate thousands of distinct mRNA transcripts and protein isoforms. The majority of transcripts are produced from the upstream promoter and encode alpha-neurexin isoforms; a much smaller number of transcripts are produced from the downstream promoter and encode beta-neurexin isoforms. The alpha-neurexins contain epidermal growth factor-like (EGF-like) sequences and laminin G domains, and have been shown to interact with neurexophilins. The beta-neurexins lack EGF-like sequences and contain fewer laminin G domains than alpha-neurexins.

BBS2	0.0031785	INTRON	N/A	Defects in BBS2 are the cause of Bardet-Biedl syndrome type 2 (BBS2) [MIM:209900]. Bardet-Biedl syndrome (BBS) is a genetically heterogeneous, autosomal recessive disorder characterized by usually severe pigmentary retinopathy, early onset obesity, polydactyly, hypogenitalism, renal malformation and mental retardation. Secondary features include diabetes mellitus, hypertension and congenital heart disease. A relatively high incidence of BBS is found in the mixed Arab populations of Kuwait and in Bedouin tribes throughout the Middle East, most likely due to the high rate of consaguinity in these populations and a founder effect.
IL1RAPL1	0.0032566	INTRON	INTRON	The protein encoded by this gene is a member of the interleukin 1 receptor family and is similar to the interleukin 1 accessory proteins. It is most closely related to interleukin 1 receptor accessory protein-like 2 (IL1RAPL2). This gene and IL1RAPL2 are located at a region on chromosome X that is associated with X-linked non-syndromic mental retardation. Deletions and mutations in this gene were found in patients with mental retardation. This gene is expressed at a high level in post-natal brain structures involved in the hippocampal memory system, which suggests a specialized role in the physiological processes underlying memory and learning abilities.
PPFIA2	0.0032884	INTRON	INTRON	The protein encoded by this gene is a member of the LAR protein-tyrosine phosphatase-interacting protein (liprin) family. Liprins interact with members of LAR family of transmembrane protein tyrosine phosphatases, which are known to be important for axon guidance and mammary gland development. It has been proposed that liprins are multivalent proteins that form complex structures and act as scaffolds for the recruitment and anchoring of LAR family of tyrosine phosphatases. This protein is most closely related to PPFIA1, a

				liprin family member known to interact with the protein phosphatase LAR. The expression of this gene is found to be downregulated by androgens in a prostate cancer cell line. Expressed only in brain
ARNTL	0.0042864	INTRON	INTRON	Brain-muscle-ARNT-like transcription factor 2d
KCNJ6	0.0046492	INTRON	INTRON	This potassium channel may be involved in the regulation of insulin secretion by glucose and/or neurotransmitters acting through G-protein-coupled receptors. Inward rectifier potassium channels are characterized by a greater tendancy to allow potassium to flow into the cell rather than out of it. Their voltage dependence is regulated by the concentration of extracellular potassium; as external potassium is raised, the voltage range of the channel opening shifts to more positive voltages. The inward rectification is mainly due to the blockage of outward current by internal magnesium. Most abundant in cerebellum, and to a lesser degree in islets and exocrine pancreas.
LMO3	0.0046674	INTRON	N/A	LIM-only protein 3 (Neuronal-specific transcription factor DAT1)
SH3GL3	0.0055678	INTRON	INTRON	May play a regulatory role in synaptic vesicle recycling. Brain and testis.
PPFIA2	0.0063903	INTRON	INTRON	The protein encoded by this gene is a member of the LAR protein-tyrosine phosphatase-interacting protein (liprin) family. Liprins interact with members of LAR family of transmembrane protein tyrosine phosphatases, which are known to be important for axon guidance and mammary gland development. It has been proposed that liprins are multivalent proteins that form complex structures and act as scaffolds for the recruitment and anchoring of LAR family of tyrosine phosphatases. This protein is most closely related to PPFIA1, a liprin family member known to interact with the protein phosphatase LAR. The expression of this gene is found to be downregulated by androgens in a prostate cancer cell line.

NTRK2	0.0070383	INTRON	INTRON	Receptor for brain-derived neurotrophic factor (BDNF), neurotrophin-3 and neurotrophin-4/5 but not nerve growth factor (NGF). Involved in the development and/or maintenance of the nervous system. This is a tyrosine-protein kinase receptor. Known substrates for the TRK receptors are SHC1, PI-3 kinase, and PLC- gamma-1.Isoform TrkB is widely expressed, mainly in the nervous tissue. In the CNS, expression is observed in the cerebral cortex, hippocampus, thalamus, choroid plexus, granular layer of the cerebellum, brain stem, and spinal cord. In the peripheral nervous system, it is expressed in many cranial ganglia, the ophtalmic nerve, the vestibular system, multiple facial structures, the submaxillary glands, and dorsal root ganglia. Isoform TrkB-T1 is expressed in multiple tissues, mainly in brain, pancreas, kidney and heart. Isoform TrkB-T-Shc is predominantly expressed in brain.
NAV3	0.0075114	INTRON	INTRON	This gene belongs to the neuron navigator family and is expressed predominantly in the nervous system. The encoded protein contains coiled-coil domains and a conserved AAA domain characteristic for ATPases associated with a variety of cellular activities.
GRIA4	0.0085929	INTRON	INTRON	Human glutamate receptor 4 (GRIA4) is a new member of the family of ionotropic glutamate receptors which are the predominant excitatory neurotransmitter receptors in the mammalian brain. Binding studies showed that human GRIA4 transfected into simian kidney cells (COS-1) exhibits high specific binding for[3H](RS)-alpha-amino- 3-hydroxy-5-methylisoxazole-4-propionic acid ([3H]AMPA) but not [3H]kainate. Ion substitution experiments indicate that hGluR4 receptor-linked ion channels in their homomeric state are permeable to both CA2+ and Na+ ions. Immunoprecipitation studies suggest that GRIA1 exists in situ in

				the form of a pentamer.
SRGAP3	0.0098257	INTRON	INTRON	GTPase-activating protein for RAC1 and perhaps Cdc42, but not for RhoA small GTPase. May attenuate RAC1 signaling in neurons. Highly expressed in adult and fetal brain. Expressed at low levels in kidney.
CNTNAP4	2.47E-05	N/A	INTRON	This gene product belongs to the neurexin family, members of which function in the vertebrate nervous system as cell adhesion molecules and receptors. This protein, like other neurexin proteins, contains epidermal growth factor repeats and laminin G domains. In addition, it includes an F5/8 type C domain, discoidin/neuropilin- and fibrinogen-like domains, and thrombospondin N-terminal-like domains. Alternative splicing results in two transcript variants encoding different isoforms.
HEXB	7.60E-05	N/A	INTRON	Hexosaminidase B is the beta subunit of the lysosomal enzyme beta-hexosaminidase that, together with the cofactor GM2 activator protein, catalyzes the degradation of the ganglioside GM2, and other molecules containing terminal N-acetyl hexosamines. Beta-hexosaminidase is composed of two subunits, alpha and beta, which are encoded by separate genes. Both beta-hexosaminidase alpha and beta subunits are members of family 20 of glycosyl hydrolases. Mutations in the alpha or beta subunit genes lead to an accumulation of GM2 ganglioside in neurons and neurodegenerative disorders termed the GM2 gangliosidoses. Beta subunit gene mutations lead to Sandhoff disease (GM2-gangliosidosis type II).
NRP2	0.0001402	N/A	INTRON	This gene encodes a member of the neuropilin family of receptor proteins. The encoded transmembrane protein binds to SEMA3C protein {sema domain, immunoglobulin domain (Ig), short basic domain, secreted,

				(semaphorin) 3C} and SEMA3F protein {sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3F}, and interacts with vascular endothelial growth factor (VEGF). This protein may play a role in cardiovascular development, axon guidance, and tumorigenesis. Multiple transcript variants encoding distinct isoforms have been identified for this gene.
UNC5A	0.0002723	N/A	INTRON	Receptor for netrin required for axon guidance. Mediates axon repulsion of neuronal growth cones in the developing nervous system upon ligand binding. Axon repulsion in growth cones may be caused by its association with DCC that may trigger signaling for repulsion. It also acts as a dependence receptor required for apoptosis induction when not associated with netrin ligand
SLIT1	0.000352	N/A	INTRON	Thought to act as molecular guidance cue in cellular migration, and function appears to be mediated by interaction with roundabaout homolog receptors. During neural development involved in axonal navigation at the ventral midline of the neural tube and projection of axons to different regions (By similarity). SLIT1 and SLIT2 together seem to be essential for midline guidance in the forebrain by acting as repulsive signal preventing inappropriate midline crossing by axons projecting from the olfactory bulb. Predominantly expressed in adult forebrain. Expressed in fetal brain, lung and kidney.
SEMA4F	0.0004059	N/A	UPSTREAM	Has growth cone collapse activity against retinal ganglion-cell axons (By similarity).
SNX10	0.0006223	N/A	UPSTREAM	This gene encodes a member of the sorting nexin family. Members of this family contain a phox (PX) domain, which is a phosphoinositide binding domain, and are involved in intracellular trafficking. This protein does not contain a coiled coil region, like some family members. This gene encodes a protein whose function has not been determined.

OCA2	0.0006815	N/A	INTRON	OCA2 encodes the human homologue of the mouse p (pink-eyed dilution) gene. The P protein is believed to be an integral membrane protein involved in small molecule transport, specifically tyrosine - a precursor of melanin. Mutations in OCA2 result in type 2 oculocutaneous albinism.
D4S234E	0.0006999	N/A	INTRON	Neuron-specific protein family member 1 (Brain neuron cytoplasmic protein 1) (D4S234).
GABRG2	0.0008157	N/A	INTRON	Gamma-aminobutyric acid (GABA), the major inhibitory neurotransmitter in the brain, mediates neuronal inhibition by binding to GABA receptors. The type A GABA receptors are pentameric chloride channels assembled from among many genetic variants of GABA(A) subunits. This gene encodes the gamma 2 subunit of GABA(A) receptor. Mutations in this gene have been associated with epilepsy and febrile seizures. Alternative splicing of this gene results in transcript variants encoding different isoforms.
RCOR1	0.000956	N/A	INTRON	The RCOR gene encodes a functional corepressor required for regulation of neural-specific gene expression.[supplied by OMIM]. Ubiquitously expressed in adult tissues.
CRSP2	0.0010282	N/A	INTRON	The activation of gene transcription is a multistep process that is triggered by factors that recognize transcriptional enhancer sites in DNA. These factors work with co-activators to direct transcriptional initiation by the RNA polymerase II apparatus. The protein encoded by this gene is a subunit of the CRSP (cofactor required for SP1 activation) complex, which, along with TFIID, is required for efficient activation by SP1. This protein is also a component of other multisubunit complexes e.g. thyroid hormone receptor-(TR-) associated proteins which interact with TR and facilitate TR function on DNA templates in conjunction with initiation factors and cofactors. This protein contains a bipartite nuclear

				localization signal. This gene is known to escape chromosome X-inactivation.
PTPRR	0.0010707	N/A	INTRON	The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. This PTP possesses an extracellular region, a single transmembrane region, and a single intracellular catalytic domains, and thus represents a receptor-type PTP. The similar gene predominately expressed in mouse brain was found to associate with, and thus regulate the activity and cellular localization of MAP kinases. The rat counterpart of this gene was reported to be regulated by the nerve growth factor, which suggested the function of this gene in neuronal growth and differentiation.
THRAP3	0.0020327	N/A	INTRON	Plays a role in transcriptional coactivation. Ubiquitious
SNX24	0.0023386	N/A	INTRON	May be involved in several stages of intracellular trafficking (By similarity).
CNDP1	0.003322	N/A	INTRON	This gene encodes a member of the M20 metalloprotease family. The encoded protein is specifically expressed in the brain, is a homodimeric dipeptidase which was identified as human carnosinase. This gene contains trinucleotide (CTG repeat length polymorphism in the coding region.
IKBKAP	0.0040589	N/A	INTRON	The protein encoded by this gene is a scaffold protein and a regulator for 3 different kinases involved in proinflammatory signaling. This encoded protein can bind NF-kappa-B-inducing kinase (NIK) and IKKs through separate domains and assemble them into an active kinase complex. Mutations in this gene have been associated with familial dysautonomia.
CORO2B	0.0043873	N/A	INTRON	May play a role in the reorganization of neuronal actin structure. Expressed predominantly

				in brain.
IQSEC3	0.0047263	N/A	INTRON	IQ motif and Sec7 domain 3
WASF1	0.0052691	N/A	INTRON	The protein encoded by this gene, a member of the Wiskott-Aldrich syndrome protein (WASP)-family, plays a critical role downstream of Rac, a Rho-family small GTPase, in regulating the actin cytoskeleton required for membrane ruffling. associate with an actin nucleation core Arp2/3 complex while enhancing actin polymerization in vitro. Wiskott-Aldrich syndrome is a disease of the immune system, likely due to defects in regulation of actin cytoskeleton. Multiple alternatively spliced transcript variants encoding the same protein have been found for this gene.
BPNT1	0.0057783	N/A	INTRON	BPNT1, also called bisphosphate 3-prime-nucleotidase, or BPntase, is a member of a magnesium-dependent phosphomonoesterase family. Lithium, a major drug used to treat manic depression, acts as an uncompetitive inhibitor of BPntase. The predicted human protein is 92% identical to mouse BPntase. BPntase's physiologic role in nucleotide metabolism may be regulated by inositol signaling pathways. The inhibition of human BPntase may account for lithium-induced nephrotoxicity.
PEX5L	0.006562	N/A	INTRON	PEX5-related protein (Peroxin-5-related protein) (Pex5Rp) (PEX5-like protein) (PEX2-related protein). Mainly expressed in brain. Also expressed in pancreas, testis and pituitary.
ZNF483	0.0069367	N/A	INTRON	Zinc finger protein 483. May function as a transcription factor
NPC1	0.0084211	N/A	INTRON	NPC1 was identified as the gene that when mutated, results in Niemann-Pick C disease. NPC1 encodes a putative integral membrane protein containing motifs consistent with a role in intracellular transport of cholesterol to post-lysosomal destinations.

CACNB4	0.0096159	N/A	INTRON	This gene encodes a member of the beta subunit family, a protein in the voltage-dependent calcium channel complex. Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization and consist of a complex of alpha-1, alpha-2/delta, beta, and gamma subunits in a 1:1:1:1 ratio. Various versions of each of these subunits exist, either expressed from similar genes or the result of alternative splicing. The protein described in this record plays an important role in calcium channel function by modulating G protein inhibition, increasing peak calcium current, controlling the alpha-1 subunit membrane targeting and shifting the voltage dependence of activation and inactivation. Certain mutations in this gene have been associated with idiopathic generalized epilepsy (IGE) and juvenile myoclonic epilepsy (JME). Alternate transcriptional splice variants of this gene, encoding different isoforms, have been characterized.

Table B.2: List of selected genes significantly differentially expressed in testis and associated with retrotransposon INDEL variation

Gene Symbol	p-value of differential expression	Human Gap Location	Chimp Gap Location	Gene Function
PBK	4.70E-08	INTRON	INTRON	The protein encoded by this gene is a serine/threonine kinase related to the dual specific mitogen-activated protein kinase kinase (MAPKK) family. Evidence suggests that mitotic phosphorylation is required for its catalytic activity. This mitotic kinase may be involved in the activation of lymphoid cells and support testicular functions, with a suggested role in the process of spermatogenesis. Expressed in the testis and placenta. In the testis, restrictedly expressed in outer cell layer of seminiferous tubules.

TUBA2	9.35E-08	UPSTREAM	N/A	Microtubules of the eukaryotic cytoskeleton perform essential and diverse functions and are composed of a heterodimer of alpha and beta tubulin. The genes encoding these microtubule constituents are part of the tubulin superfamily, which is composed of six distinct families. Genes from the alpha, beta and gamma tubulin families are found in all eukaryotes. The alpha and beta tubulins represent the major components of microtubules, while gamma tubulin plays a critical role in the nucleation of microtubule assembly. There are multiple alpha and beta tubulin genes and they are highly conserved among and between species. This gene is an alpha tubulin gene that encodes a protein identical to the mouse testis-specific Tuba3 and Tuba7 gene products. This gene is located in the 13q11 region, which is associated with the genetic diseases Clouston hidrotic ectodermal dysplasia and Kabuki syndrome. Alternative splicing has been observed for this gene and two variants have been identified.
CATSPER3	1.97E-06	INTRON	N/A	cation channel, sperm associated 3
ATR	4.17E-06	INTRON	INTRON	The protein encoded by this gene belongs the PI3/PI4-kinase family, and is most closely related to ATM, a protein kinase encoded by the gene mutated in ataxia telangiectasia. This protein and ATM share similarity with Schizosaccharomyces pombe rad3, a cell cycle checkpoint gene required for cell cycle arrest and DNA damage repair in response to DNA damage. This kinase has been shown to phosphorylate checkpoint kinase CHK1, checkpoint proteins RAD17, and RAD9, as well as tumor suppressor protein BRCA1. Mutations of this gene are associated with Seckel syndrome. An alternatively spliced transcript variant of this gene has been reported, however, its full length nature is not known. Transcript variants utilizing alternative polyA sites exist.
DDX27	6.56E-06	INTRON	INTRON	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. This gene encodes a DEAD box protein, the function of which has not been determined.

TESK2	2.09E-05	INTRON	N/A	This gene product is a serine/threonine protein kinase that contains an N-terminal protein kinase domain that is structurally similar to the kinase domains of testis-specific protein kinase-1 and the LIM motif-containing protein kinases (LIMKs). Its overall structure is most related to the former, indicating that it belongs to the TESK subgroup of the LIMK/TESK family of protein kinases. This gene is predominantly expressed in testis and prostate. The developmental expression pattern of the rat gene in testis suggests an important role for this gene in meitoic stages and/or early stages of spermiogenesis. Predominantly expressed in testis and prostate. Found predominantly in nongerminal Sertoli cells.
GALNS	2.13E-05	INTRON	INTRON	This gene encodes N-acetylgalactosamine-6-sulfatase which is a lysosomal exohydrolase required for the degradation of the glycosaminoglycans, keratan sulfate, and chondroitin 6-sulfate. Sequence alterations including point, missense and nonsense mutations, as well as those that affect splicing, result in a deficiency of this enzyme. Deficiencies of this enzyme lead to Morquio A syndrome, a lysosomal storage disorder. Defects in GALNS are the cause of mucopolysaccharidosis type IVA (MPS-IVA) [MIM:253000]; also known as Morquio A syndrome. MPS-IVA is characterized by specific spondyloepiphyseal dysplasia, short trunk dwarfism, coxa valga, odontoid hypoplasia, corneal opacities, preservation of intelligence, and excessive urinary excretion of keratan sulfate and chondroitin-6-sulfate. Severely affected patients usually die of cardiopulmonary disturbance or cervical cord compression in the second or third decade of life.
PDE10A	2.50E-05	INTRON	INTRON	Plays a role in signal transduction by regulating the intracellular concentration of cyclic nucleotides. This enzyme can hydrolyze both cAMP and cGMP, having a higher affinity for cAMP. Abundant in the putamen and caudate nucleus regions of brain and testis, moderately expressed in the thyroid gland, pituitary gland, thalamus and cerebellum.
DYM	4.67E-05	INTRON	INTRON	This gene encodes a protein which is necessary for normal skeletal development and brain function. Mutations in this gene are associated with two types of recessive osteochondrodysplasia, Dyggve-Melchior-Clausen (DMC) dysplasia and Smith-McCort (SMC) dysplasia, which involve both skeletal defects and mental retardation. Defects in DYM are the cause of Dyggve-Melchior-Clausen syndrome (DMC) [MIM:223800]. DMC is a rare autosomal recessive disorder characterized by short trunk dwarfism, microcephaly and psychomotor retardation. Electron microscopic study of cutaneous cells of affected patients shows dilated rough endoplasmic reticulum, enlarged and aberrant vacuoles and numerous vesicles. DMC is progressive. Defects in DYM are the cause of Smith-McCort

				dysplasia (SMC) [MIM:607326]. SMC is a rare autosomal recessive osteochondrodysplasia characterized by short limbs and trunk with barrel-shaped chest. The radiographic phenotype includes platyspondyly, generalized abnormalities of the epiphyses and metaphyses, and a distinctive lacy appearance of the iliac crest, features identical to those of Dyggve-Melchior-Clausen syndrome.
DDX4	6.86E-05	INTRON	INTRON	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. This gene encodes a DEAD box protein, which is a homolog of VASA proteins in Drosophila and several other species. The gene is specifically expressed in the germ cell lineage in both sexes and functions in germ cell development.
LPP	7.99E-05	INTRON	INTRON	May play a structural role at sites of cell adhesion in maintaining cell shape and motility. In addition to these structural functions, it may also be implicated in signaling events and activation of gene transcription. May be involved in signal transduction from cell adhesion sites to the nucleus allowing successful integration of signals arising from soluble factors and cell-cell adhesion sites. Also suggested to serve as a scaffold protein upon which distinct protein complexes are assembled in the cytoplasm and in the nucleus.
USP9Y	9.42E-05	INTRON	INTRON	This gene is a member of the peptidase C19 family. It encodes a protein that is similar to ubiquitin-specific proteases, which cleave the ubiquitin moiety from ubiquitin-fused precursors and ubiquitinylated proteins. Mutations in this gene have been associated with Sertoli cell-only (SCO) syndrome and male infertility.
CDKL5	9.89E-05	INTRON	INTRON	This gene is a member of Ser/Thr protein kinase family and encodes a phosphorylated protein with protein kinase activity. Mutations in this gene have been associated with X-linked infantile spasm syndrome (ISSX), also known as X-linked West syndrome, and Rett syndrome (RTT). Alternate transcriptional splice variants have been characterized.

DMRT1	0.00010	INTRON	INTRON	This gene is found in a cluster with two other members of the gene family, having in common a zinc finger-like DNA-binding motif (DM domain). The DM domain is an ancient, conserved component of the vertebrate sex-determining pathway that is also a key regulator of male development in flies and nematodes. This gene exhibits a gonad-specific and sexually dimorphic expression pattern. Defective testicular development and XY feminization occur when this gene is hemizygous.
FANCE	0.00013	INTRON	N/A	The Fanconi anemia complementation group (FANC) currently includes FANCA, FANCB, FANCC, FANCD1 (also called BRCA2), FANCD2, FANCE, FANCF, FANCG, and FANCL. Fanconi anemia is a genetically heterogeneous recessive disorder characterized by cytogenetic instability, hypersensitivity to DNA crosslinking agents, increased chromosomal breakage, and defective DNA repair. The members of the Fanconi anemia complementation group do not share sequence similarity; they are related by their assembly into a common nuclear protein complex. This gene encodes the protein for complementation group E.
DHX34	0.00015	INTRON	INTRON	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this DEAD box protein family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. This gene encodes a member of this family. It is mapped to the glioma 19q tumor suppressor region and is a tumor suppressor candidate gene. Two alternative transcripts encoding different isoforms have been described.
FANCD2	0.00024	EXON INTRON	INTRON	The Fanconi anemia complementation group (FANC) currently includes FANCA, FANCB, FANCC, FANCD1 (also called BRCA2), FANCD2, FANCE, FANCF, FANCG, and FANCL. Fanconi anemia is a genetically heterogeneous recessive disorder characterized by cytogenetic instability, hypersensitivity to DNA crosslinking agents, increased chromosomal breakage, and defective DNA repair. The members of the Fanconi anemia complementation group do not share sequence similarity; they are related by their assembly into a common nuclear protein complex. This gene encodes the protein for complementation group D2. This protein is monoubiquinated in response to DNA damage, resulting in its localization to nuclear foci with other proteins (BRCA1 AND BRCA2) involved in homology-directed DNA repair. Alternative splicing results in two transcript variants encoding

				different isoforms. Highly expressed in germinal center cells of the spleen, tonsil, and reactive lymph nodes, and in the proliferating basal layer of squamous epithelium of tonsil, esophagus, oropharynx, larynx and cervix. Expressed in cytotrophoblastic cells of the placenta and exocrine cells of the pancreas (at protein level). Highly expressed in testis, where expression is restricted to maturing spermatocytes.
GPC5	0.00025	INTRON	INTRON	Cell surface heparan sulfate proteoglycans are composed of a membrane-associated protein core substituted with a variable number of heparan sulfate chains. Members of the glypican-related integral membrane proteoglycan family (GRIPS) contain a core protein anchored to the cytoplasmic membrane via a glycosyl phosphatidylinositol linkage. These proteins may play a role in the control of cell division and growth regulation.
GABPA	0.00049	INTRON	N/A	This gene encodes one of three GA-binding protein transcription factor subunits which functions as a DNA-binding subunit. Since this subunit shares identity with a subunit encoding the nuclear respiratory factor 2 gene, it is likely involved in activation of cytochrome oxidase expression and nuclear control of mitochondrial function. This subunit also shares identity with a subunit constituting the transcription factor E4TF1, responsible for expression of the adenovirus E4 gene. Because of its chromosomal localization and ability to form heterodimers with other polypeptides, this gene may play a role in the Down Syndrome phenotype.
GAS7	0.00049	INTRON	INTRON	Growth arrest-specific 7 is expressed primarily in terminally differentiated brain cells and predominantly in mature cerebellar Purkinje neurons. GAS7 plays a putative role in neuronal development. Several transcript variants encoding proteins which vary in the N-terminus have been described. May play a role in promoting maturation and morphological differentiation of cerebellar neurons.
ZFYVE9	0.00058	INTRON	N/A	This gene encodes a double zinc finger (FYVE domain) protein that interacts directly with SMAD2 and SMAD3, and is involved in Alzheimer's disease. SMAD proteins transmit signals from transmembrane serine/threonine kinase receptors to the nucleus. The FYVE domain has been identified in a number of unrelated signaling molecules. This protein functions to recruit SMAD2 to the transforming growth factor-beta receptor. The FYVE domain is required to maintain the normal localization of this protein but is not involved in mediating interaction with SMADs.

				The C-terminal domain of this protein interacts with the TGFB receptor. This protein is a component of the TGFB pathway that brings the SMAD substrate to the receptor. Three alternatively spliced transcripts encoding distinct isoforms have been found for this gene.
ROBO1	0.00078	INTRON	INTRON	Bilateral symmetric nervous systems have special midline structures that establish a partition between the two mirror image halves. Some axons project toward and across the midline in response to long-range chemoattractants emanating from the midline. In Drosophila, the roundabout gene, a member of the immunoglobulin gene superfamily, encodes an integral membrane protein that is both an axon guidance receptor and a cell adhesion receptor. This receptor is involved in the decision by axons to cross the central nervous system midline. The protein encoded by this gene is structurally similar to the Drosophila roundabout protein. Two transcript variants encoding different isoforms have been found for this gene.
EYA2	0.00098	INTRON	INTRON	This gene encodes a member of the eyes absent (EYA) family of proteins. The encoded protein may be post-translationally modified and may play a role in eye development. A similar protein in mice can act as a transcriptional activator. Five transcript variants encoding three distinct isoforms have been identified for this gene.
HTR7	0.00099	INTRON	N/A	The neurotransmitter, serotonin, is thought to play a role in various cognitive and behavioral functions. The serotonin receptor encoded by this gene belongs to the superfamily of G protein-coupled receptors and the gene is a candidate locus for involvement in autistic disorder and other neuropsychiatric disorders. Three splice variants have been identified which encode proteins that differ in the length of their carboxy terminal ends.
EYA4	0.00100	INTRON	INTRON	This gene encodes a member of the eyes absent (EYA) family of proteins. The encoded protein may act as a transciptional activator and be important for continued function of the mature organ of Corti. Mutations in this gene are associated with postlingual, progressive, autosomal dominant hearing loss at the deafness, autosomal dominant nonsyndromic sensorineural 10 locus. Three transcript variants encoding distinct isoforms have been identified for this gene.
DNM3	0.00108	INTRON	INTRON	Dynamin-3 (EC 3.6.5.5) (Dynamin, testicular) (T-dynamin).Microtubule-associated force-producing protein involved in producing microtubule bundles and able to bind and hydrolyze GTP. Most probably involved in vesicular trafficking processes, in

				particular endocytosis
SPACA4	0.00134	DOWNSTREAM	N/A	sperm acrosomal membrane protein 14
DUSP5	0.00151	UPSTREAM INTRON		The protein encoded by this gene is a member of the dual specificity protein phosphatase subfamily. These phosphatases inactivate their target kinases by dephosphorylating both the phosphoserine/threonine and phosphotyrosine residues. They negatively regulate members of the mitogen-activated protein (MAP) kinase superfamily (MAPK/ERK, SAPK/JNK, p38), which are associated with cellular proliferation and differentiation. Different members of the family of dual specificity phosphatases show distinct substrate specificities for various MAP kinases, different tissue distribution and subcellular localization, and different modes of inducibility of their expression by extracellular stimuli. This gene product inactivates ERK1, is expressed in a variety of tissues with the highest levels in pancreas and brain, and is localized in the nucleus.
GPR64	0.00174	INTRON	INTRON	Could be involved in a signal transduction pathway controlling epidymal function and male fertility.
CTCFL	0.00208	INTRON	N/A	CCCTC-binding factor (CTCF), an 11-zinc-finger factor involved in gene regulation, utilizes different zinc fingers to bind varying DNA target sites. CTCF forms methylation-sensitive insulators that regulate X-chromosome inactivation. This gene is a paralog of CTCF and appears to be expressed primarily in the cytoplasm of spermatocytes, unlike CTCF which is expressed primarily in the nucleus of somatic cells. CTCF and the protein encoded by this gene are normally expressed in a mutually exclusive pattern that correlates with resetting of methylation marks during male germ cell differentiation.
RUNX2	0.00217	INTRON	N/A	This gene is a member of the RUNX family of transcription factors and encodes a nuclear protein with an Runt DNA-binding domain. This protein is essential for osteoblastic differentiation and skeletal morphogenesis, acting as a scaffold for nucleic acids and regulatory factors involved in skeletal gene expression. The protein can bind DNA both as a monomer or, with more affinity, as a subunit of a heterodimeric complex. Mutations in this gene have been associated with the bone development disorder cleidocranial dysplasia (CCD). Transcript variants, encoding different protein isoforms, result from alternate promoter use as well as alternate splicing.

DUSP12	0.00233	INTRON	N/A	The protein encoded by this gene is a member of the dual specificity protein phosphatase subfamily. These phosphatases inactivate their target kinases by dephosphorylating both the phosphoserine/threonine and phosphotyrosine residues. They negatively regulate members of the mitogen-activated protein (MAP) kinase superfamily (MAPK/ERK, SAPK/JNK, p38), which is associated with cellular proliferation and differentiation. Different members of the family of dual specificity phosphatases show distinct substrate specificities for various MAP kinases, different tissue distribution and subcellular localization, and different modes of inducibility of their expression by extracellular stimuli. This gene product is the human ortholog of the Saccharomyces cerevisiae YVH1 protein tyrosine phosphatase. It is localized predominantly in the nucleus, and is novel in that it contains, and is regulated by a zinc finger domain. Ubiquitous, highest expression in spleen, testis, ovary, and peripheral blood leukocytes and lower expression in liver and lung
RAB3GAP2	0.00234	INTRON	INTRON	Regulatory subunit of a GTPase activating protein that has specificity for Rab3 subfamily (RAB3A, RAB3B, RAB3C and RAB3D). Rab3 proteins are involved in regulated exocytosis of neurotransmitters and hormones. Rab3 GTPase-activating complex specifically converts active Rab3-GTP to the inactive form Rab3- GDP. Required for normal eye and brain development. May participate in neurodevelopmental processes such as proliferation, migration and differentiation before synapse formation, and nonsynaptic vesicular release of neurotransmitters.
RNPS1	0.00253	INTRON	N/A	This gene encodes a protein that is part of a post-splicing multiprotein complex involved in both mRNA nuclear export and mRNA surveillance. mRNA surveillance detects exported mRNAs with truncated open reading frames and initiates nonsense-mediated mRNA decay (NMD). When translation ends upstream from the last exon-exon junction, this triggers NMD to degrade mRNAs containing premature stop codons. This protein binds to the mRNA and remains bound after nuclear export, acting as a nucleocytoplasmic shuttling protein. This protein contains many serine residues. Two splice variants have been found for this gene; both variants encode the same protein.
SLC25A15	0.00287	INTRON	N/A	Defects in SLC25A15 are the cause of hyperornithinemia- hyperammonemia-homocitrullinuria syndrome (HHH syndrome) [MIM:238970]. It is an autosomal recessive disorder resulting in various neurologic symptoms, including mental retardation, spastic paraparesis with pyramidal signs, cerebellar ataxia, and episodic disturbance of consciousness or coma caused by hyperammonemia. It causes a functional impairment of the urea cycle.

DIAPH2	0.00366	INTRON	INTRON	This gene may play a role in the development and normal function of the ovaries. Mutations of this gene have been linked to premature ovarian failure. Alternative splicing results in two protein isoforms.
RAI1	0.00411	INTRON	INTRON	This gene is located within the Smith-Magenis syndrome region on chromosome 17. It is highly similar to its mouse counterpart and is expressed at high levels mainly in neuronal tissues. The protein encoded by this gene includes a polymorphic polyglutamine tract in the N-terminal domain. Expression of the mouse counterpart in neurons is induced by retinoic acid. This gene is associated with both the severity of the phenotype and the response to medication in schizophrenic patients.
RFC1	0.00575	INTRON	INTRON	The protein encoded by this gene is the large subunit of replication factor C, which is a five subunit DNA polymerase accessory protein. Replication factor C is a DNA-dependent ATPase that is required for eukaryotic DNA replication and repair. The protein acts as an activator of DNA polymerases, binds to the 3' end of primers, and promotes coordinated synthesis of both strands. It also may have a role in telomere stability.
MCM5	0.00650	INTRON	N/A	The protein encoded by this gene is structurally very similar to the CDC46 protein from S. cerevisiae, a protein involved in the initiation of DNA replication. The encoded protein is a member of the MCM family of chromatin-binding proteins and can interact with at least two other members of this family. The encoded protein is upregulated in the transition from the G0 to G1/S phase of the cell cycle and may actively participate in cell cycle regulation.
RP2	0.00661	INTRON	INTRON	The RP2 locus has been implicated as one cause of X-linked retinitis pigmentosa. The predicted gene product shows homology with human cofactor C, a protein involved in the ultimate step of beta-tubulin folding. Progressive retinal degeneration may therefore be due to the accumulation of incorrectly-folded photoreceptor or neuron-specific tubulin isoforms followed by progressive cell death.
TBX1	0.00804	INTRON	N/A	This gene is a member of a phylogenetically conserved family of genes that share a common DNA-binding domain, the T-box. T-box genes encode transcription factors involved in the regulation of developmental processes. This gene product shares 98% amino acid sequence identity with the mouse ortholog. DiGeorge syndrome (DGS)/velocardiofacial syndrome (VCFS), a common congenital disorder characterized by neural-crest-related developmental defects, has been associated with deletions of chromosome 22q11.2, where this gene has been mapped. Studies using mouse models of DiGeorge syndrome suggest a major role for this gene in the molecular etiology of DGS/VCFS. Several alternatively spliced transcript variants encoding different isoforms have been

				described for this gene.
SH3TC2	0.00805	INTRON	INTRON	This gene encodes a protein with two N-terminal Src homology 3 (SH3) domains and 10 tetratricopeptide repeat (TPR) motifs, and is a member of a small gene family. The gene product has been proposed to be an adapter or docking molecule. Mutations in this gene result in autosomal recessive Charcot-Marie-Tooth disease type 4C, a childhood-onset neurodegenerative disease characterized by demyelination of motor and sensory neurons.
PAX3	0.00815	INTRON	N/A	This gene is a member of the paired box (PAX) family of transcription factors. Members of the PAX family typically contain a paired box domain and a paired-type homeodomain. These genes play critical roles during fetal development. Mutations in paired box gene 3 are associated with Waardenburg syndrome, craniofacial-deafness-hand syndrome, and alveolar rhabdomyosarcoma. The translocation t(2;13)(q35;q14), which represents a fusion between PAX3 and the forkhead gene, is a frequent finding in alveolar rhabdomyosarcoma. Alternative splicing results in transcripts encoding isoforms with different C-termini.
POT1	0.00855	INTRON	INTRON	This gene is a member of the telombin family and encodes a nuclear protein involved in telomere maintenance. Specifically, this protein functions as a member of a multi-protein complex that binds to the TTAGGG repeats of telomeres, regulating telomere length and protecting chromosome ends from illegitimate recombination, catastrophic chromosome instability, and abnormal chromosome segregation. Increased transcriptional expression of this gene is associated with stomach carcinogenesis and its progression. Alternatively spliced transcript variants have been described.
SNCAIP	0.00890	INTRON	INTRON	This gene encodes a protein containing several protein-protein interaction domains, including ankyrin-like repeats, a coiled-coil domain, and an ATP/GTP-binding motif. The encoded protein interacts with alpha-synuclein in neuronal tissue and may play a role in the formation of cytoplasmic inclusions and neurodegeneration. A mutation in this gene has been associated with Parkinson's disease. Alternatively spliced transcript variants encoding different isoforms of this gene have been described, but their full-length nature has yet to be determined.

TOP1	0.00931	INTRON	N/A	This gene encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This enzyme catalyzes the transient breaking and rejoining of a single strand of DNA which allows the strands to pass through one another, thus altering the topology of DNA. This gene is localized to chromosome 20 and has pseudogenes which reside on chromosomes 1 and 22.
DHX33	8.84E-07	N/A	INTRON	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this DEAD box protein family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. This gene encodes a member of this family. The function of this member has not been determined.
NUP88	1.26E-06	N/A	INTRON	The nuclear pore complex is a massive structure that extends across the nuclear envelope, forming a gateway that regulates the flow of macromolecules between the nucleus and the cytoplasm. Nucleoporins, a family of 50 to 100 proteins, are the main components of the nuclear pore complex in eukaryotic cells. The protein encoded by this gene belongs to the nucleoporin family and is associated with the oncogenic nucleoporin CAN/Nup214 in a dynamic subcomplex. This protein is also overexpressed in a large number of malignant neoplasms and precancerous dysplasias.
TTLL2	3.18E-06	N/A	INTRON	Tubulin tyrosine ligase-like protein 2 (Testis-specific protein NYD- TSPG).
SPSB4	4.37E-06	N/A	INTRON	SPRY domain-containing SOCS box protein SSB-4
SAMD8	6.18E-06	N/A	INTRON	sterile alpha motif domain containing 8
ACRBP	7.09E-06	N/A	INTRON	The protein encoded by this gene is similar to proacrosin binding protein sp32 precursor found in mouse, guinea pig, and pig. This protein is located in the sperm acrosome and is thought to function as a binding protein to proacrosin for packaging and condensation of the acrosin zymogen in the acrosomal matrix. This protein is a member of the cancer/testis family of antigens and it is found to be immunogenic. In normal tissues, this mRNA is expressed only in testis, whereas it is detected in a range of different tumor types such as bladder, breast, lung, liver, and colon.

MPP6	7.15E-06	N/A	INTRON	Members of the peripheral membrane-associated guanylate kinase (MAGUK) family function in tumor suppression and receptor clustering by forming multiprotein complexes containing distinct sets of transmembrane, cytoskeletal, and cytoplasmic signaling proteins. All MAGUKs contain a PDZ-SH3-GUK core and are divided into 4 subfamilies, DLG-like (see DLG1; MIM 601014), ZO1-like (see TJP1; MIM 601009), p55-like (see MPP1; MIM 305360), and LIN2-like (see CASK; MIM 300172), based on their size and the presence of additional domains. MPP6 is a member of the p55-like MAGUK subfamily (Tseng et al., 2001). Abundant in testis, brain, and kidney with lower levels detectable in other tissues.
NYD-SP26	9.35E-06	N/A	UPSTREAM	testis development protein NYD-SP26
PACRG	9.71E-06	INTRON	INTRON	Suppresses cell death induced by accumulation of unfolded Pael receptor (Pael-R, a substrate of Parkin). Facilitates the formation of inclusions consisting of Pael-R, molecular chaperones, protein degradation molecules and itself when proteasome is inhibited. May play an important role in the formation of Lewy bodies and protection of dopaminergic neurons against Parkinson disease.
PSMA8	2.97E-05	N/A	INTRON	The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH. The proteasome has an ATP-dependent proteolytic activity (By similarity). This may be a testis-specific subunit.
UCK2	6.54E-05	N/A	INTRON	The protein encoded by this gene catalyzes the phosphorylation of uridine monophosphate to uridine diphosphate. This is the first step in the production of the pyrimidine nucleoside triphosphates required for RNA and DNA synthesis. In addition, an allele of this gene may play a role in mediating nonhumoral immunity to Hemophilus influenzae type B. Testis-specific (Ref.1); found exclusively in the placenta (Ref.2).
KAL1	8.68E-05	N/A	INTRON	Mutations in the KAL1 gene cause the X-linked Kallmann syndrome. The predictated KAL1 protein sequence is similar to proteins known to function in neural cell adhesion and axonal migration.
VDAC3	9.83E-05	N/A	UPSTREAM	Forms a channel through the mitochondrial outer membrane that allows diffusion of small hydrophilic molecules. Widely expressed. Highest in testis.

CHAF1A	0.00013	N/A	INTRON	Chromatin assembly factor I (CAF1) is a nuclear complex consisting of p50, p60 (CHAF1B; MIM 601245), and p150 (CHAF1A) subunits that assembles histone octamers onto replicating DNA in vitro (Kaufman et al., 1995).[supplied by OMIM]. Complex that is thought to mediate chromatin assembly in DNA replication and DNA repair. Assembles histone octamers onto replicating DNA in vitro. CAF-1 performs the first step of the nucleosome assembly process, bringing newly synthesized histones H3 and H4 to replicating DNA; histones H2A/H2B can bind to this chromatin precursor subsequent to DNA replication to complete the histone octamer. CHAF1A binds to histones H3 and H4. It may play a role in heterochromatin maintenance in proliferating cells by bringing newly synthesized cbx proteins to heterochromatic DNA replication foci (By similarity). The CCR4-NOT complex functions as general transcription regulation complex
CDH11	0.00013	N/A	INTRON	This gene encodes a type II classical cadherin from the cadherin superfamily, integral membrane proteins that mediate calcium-dependent cell-cell adhesion. Mature cadherin proteins are composed of a large N-terminal extracellular domain, a single membrane-spanning domain, and a small, highly conserved C-terminal cytoplasmic domain. Type II (atypical) cadherins are defined based on their lack of a HAV cell adhesion recognition sequence specific to type I cadherins. Expression of this particular cadherin in osteoblastic cell lines, and its upregulation during differentiation, suggests a specific function in bone development and maintenance. Expressed mainly in brain but also found in other tissues. Expressed in neuroblasts.
SPATA3	0.00022	N/A		testis and spermatogenesis cell apoptosis
TAF4	0.00031	INTRON	INTRON	Initiation of transcription by RNA polymerase II requires the activities of more than 70 polypeptides. The protein that coordinates these activities is transcription factor IID (TFIID), which binds to the core promoter to position the polymerase properly, serves as the scaffold for assembly of the remainder of the transcription complex, and acts as a channel for regulatory signals. TFIID is composed of the TATA-binding protein (TBP) and a group of evolutionarily conserved proteins known as TBP-associated factors or TAFs. TAFs may participate in basal transcription, serve as coactivators, function in promoter recognition or modify general transcription factors (GTFs) to facilitate complex assembly and transcription initiation. This gene encodes one of the larger subunits of TFIID that has been shown to potentiate transcriptional activation by retinoic acid, thyroid hormone and vitamin D3 receptors. In addition, this subunit interacts with the transcription factor CREB, which has a glutamine-rich activation

				domain, and binds to other proteins containing glutamine-rich regions. Aberrant binding to this subunit by proteins with expanded polyglutamine regions has been suggested as one of the pathogenetic mechanisms underlying a group of neurodegenerative disorders referred to as polyglutamine diseases.
SUV39H2	0.00035	N/A	INTRON	Histone methyltransferase. Methylates Lys-9 of histone H3 and weakly histone H1 (in vitro). H3 Lys-9 methylation represents a specific tag for epigenetic transcriptional repression by recruiting HP1 proteins to methylated histones. May participate in regulation of higher order chromatin organization during spermatogenesis.
UTY	0.00038	N/A	INTRON	This gene encodes a protein containing tetratricopeptide repeats which are thought to be involved in protein-protein interactions. This protein is a minor histocompatibility antigen which may induce graft rejection of male stem cell grafts. Alternative splicing results in multiple transcript variants encoding different isofor
NSD1	0.00040	INTRON	INTRON	This gene encodes a protein containing a SET domain, 2 LXXLL motifs, 3 nuclear translocation signals (NLSs), 4 plant homeodomain (PHD) finger regions, and a proline-rich region. The encoded protein enhances androgen receptor (AR) transactivation, and this enhancement can be increased further in the presence of other androgen receptor associated coregulators. This protein may act as a nucleus-localized, basic transcriptional factor and also as a bifunctional transcriptional regulator. Mutations of this gene have been associated with Sotos syndrome and Weaver syndrome. One version of childhood acute myeloid leukemia is the result of a cryptic translocation with the breakpoints occurring within nuclear receptor-binding Su-var, enhancer of zeste, and trithorac domain protein 1 on chromosome 5 and nucleoporin, 98-kd on chromosome 11. Two transcript variants encoding distinct isoforms have been identified for this gene.
CLGN	0.00043	N/A	INTRON	Calmegin is a testis-specific endoplasmic reticulum chaperone protein. CLGN may play a role in spermatogeneisis and infertility.
NR6A1	0.00048	N/A	INTRON	This gene encodes an orphan nuclear receptor which is a member of the nuclear hormone receptor family. Its expression pattern suggests that it may be involved in neurogenesis and germ cell development. The protein can homodimerize and bind DNA, but in vivo targets have not been identified. The gene

				expresses three alternatively spliced transcript variants.
GABRG1	0.00049	N/A	INTRON	The protein encoded by this gene belongs to the ligand-gated ionic channel family. It is an integral membrane protein and plays an important role in inhibiting neurotransmission by binding to the benzodiazepine receptor and opening an integral chloride channel. This gene is clustered with three other family members on chromosome 4.
ROPN1L	0.00051	N/A	INTRON	The protein encoded by this gene is a sperm protein, which interacts with A-kinase anchoring protein, AKAP3, through the amphipathic helix region of AKAP3. Type II regulatory subunit of cAMP-dependent protein kinase (PKARII) also binds to this helix domain of AKAP3, which allows PKARII to be targeted to specific subcellular compartments. It is suggested that sperm contains several proteins that bind to AKAPs in a manner similar to PKARII, and this encoded protein may be one of them
MOV10L 1	0.00057	N/A	INTRON	This gene is similar to a mouse gene that encodes a putative RNA helicase and shows testis-specific expression. Isoform 1 is specifically expressed in testis
CEP290	0.00060	N/A	INTRON	This gene encodes a protein with 13 putative coiled-coil domains, a region with homology to SMC chromosome segregation ATPases, six KID motifs, three tropomyosin homology domains and an ATP/GTP binding site motif A. The protein is localized to the centrosome and cilia and has sites for N-glycosylation, tyrosine sulfation, phosphorylation, N-myristoylation, and amidation. Mutations in this gene have been associated with Joubert syndrome and nephronophthisis and the presence of antibodies against this protein is associated with several forms of cancer.
TBL1X	0.00082	N/A	INTRON	The protein encoded by this gene has sequence similarity with members of the WD40 repeat-containing protein family. The WD40 group is a large family of proteins, which appear to have a regulatory function. It is believed that the WD40 repeats mediate protein-protein interactions and members of the family are involved in signal transduction, RNA processing, gene regulation, vesicular trafficking, cytoskeletal assembly and may play a role in the control of cytotypic differentiation. This encoded protein is found as a subunit in corepressor SMRT (silencing mediator for retinoid and thyroid receptors) complex along with histone deacetylase 3 protein. This gene is located adjacent to the ocular albinism gene and it is thought to be involved in the pathogenesis of the ocular albinism with late-onset sensorineural deafness phenotype. This gene is highly similar to the Y chromosome TBL1Y gene.

ADAM18	0.00095	N/A	INTRON	This gene encodes a member of the ADAM (a disintegrin and metalloprotease domain) family. Members of this family are membrane-anchored proteins structurally related to snake venom disintegrins, and have been implicated in a variety of biologic processes involving cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesis. The protein encoded by this gene is a sperm surface protein. Sperm surface membrane protein that may be involved in spermatogenesis and fertilization. This is a non catalytic metalloprotease-like protein (By similarity). Expressed specifically in testis
STAU1	0.00104	N/A	INTRON	Staufen is a member of the family of double-stranded RNA (dsRNA)-binding proteins involved in the transport and/or localization of mRNAs to different subcellular compartments and/or organelles. These proteins are characterized by the presence of multiple dsRNA-binding domains which are required to bind RNAs having double-stranded secondary structures. The human homologue of staufen encoded by STAU, in addition contains a microtubule- binding domain similar to that of microtubule-associated protein 1B, and binds tubulin. The STAU gene product has been shown to be present in the cytoplasm in association with the rough endoplasmic reticulum (RER), implicating this protein in the transport of mRNA via the microtubule network to the RER, the site of translation. Five transcript variants resulting from alternative splicing of STAU gene and encoding three isoforms have been described. Three of these variants encode the same isoform, however, differ in their 5'UTR.
PCDHB3	0.00117	N/A	DOWNS TREAM	This gene is a member of the protocadherin beta gene cluster, one of three related gene clusters tandemly linked on chromosome five. The gene clusters demonstrate an unusual genomic organization similar to that of B-cell and T-cell receptor gene clusters. The beta cluster contains 16 genes and 3 pseudogenes, each encoding 6 extracellular cadherin domains and a cytoplasmic tail that deviates from others in the cadherin superfamily. The extracellular domains interact in a homophilic manner to specify differential cell-cell connections. Unlike the alpha and gamma clusters, the transcripts from these genes are made up of only one large exon, not sharing common 3' exons as expected. These neural cadherin-like cell adhesion proteins are integral plasma membrane proteins. Their specific functions are unknown but they most likely play a critical role in the establishment and function of specific cell-cell neural connections. Potential calcium-dependent cell-adhesion protein. May be involved in the establishment and maintenance of specific neuronal connections in the brain

TAF7L	0.00119	N/A	INTRON	This gene is similar to a mouse gene that encodes a TATA box binding protein-associated factor, and shows testis-specific expression.
SUHW2	0.00122	N/A	INTRON	This gene was identified by homology to other species. Its encoded protein is approximately 78-88% identical to a predicted sheep protein of unknown function. The protein is also approximately 25% identical to the Drosophila protein suppressor of hairy wing, which is a leucine zipper protein that represses the function of transcriptional enhancers of the gypsy retrotransposon.
TDP1	0.00134	N/A	INTRON	The protein encoded by this gene is involved in repairing stalled topoisomerase I-DNA complexes by catalyzing the hydrolysis of the phosphodiester bond between the tyrosine residue of topoisomerase I and the 3-prime phosphate of DNA. This protein may also remove glycolate from single-stranded DNA containing 3-prime phosphoglycolate, suggesting a role in repair of free-radical mediated DNA double-strand breaks. This gene is a member of the phospholipase D family and contains two PLD phosphodiesterase domains. Mutations in this gene are associated with the disease spinocerebellar ataxia with axonal neuropathy (SCAN1). While several transcript variants may exist for this gene, the full-length natures of only two have been described to date. These two represent the major variants of this gene and encode the same isoform.
TEX11	0.00166	N/A	INTRON	This gene is X-linked and is expressed in only male germ cells. Two alternatively spliced transcript variants encoding distinct isoforms have been found for this gene.
SIAH1	0.00182	N/A	INTRON	This gene encodes a protein that is a member of the seven in absentia homolog (SIAH) family. The protein is an E3 ligase and is involved in ubiquitination and proteasome-mediated degradation of specific proteins. The activity of this ubiquitin ligase has been implicated in the development of certain forms of Parkinson's disease, the regulation of the cellular response to hypoxia and induction of apoptosis. Alternative splicing results in several additional transcript variants, some encoding different isoforms and others that have not been fully characterized.
MAP2	0.00192	N/A	INTRON	This gene encodes a protein that belongs to the microtubule-associated protein family. The proteins of this family are thought to be involved in microtubule assembly, which is an essential step in neurogenesis. The products of similar genes in rat and mouse are neuron-specific cytoskeletal proteins that are enriched in dentrites, implicating a role in determining and stabilizing dentritic shape during neuron development. A number of alternatively spliced variants encoding distinct isforms have been described.

IL20RA	0.00212	N/A	INTRON	The protein encoded by this gene is a receptor for interleukin 20 (IL20), a cytokine that may be involved in epidermal function. The receptor of IL20 is a heterodimeric receptor complex consisting of this protein and interleukin 20 receptor beta (IL20B). This gene and IL20B are highly expressed in skin. The expression of both genes is found to be upregulated in Psoriasis. Widely expressed with highest levels in skin and testis and high levels in brain. Highly expressd in psoriatic skin.
FUCA1	0.00249	N/A	INTRON	Fucosidosis is an autosomal recessive lysosomal storage disease caused by defective alpha-L-fucosidase with accumulation of fucose in the tissues. Different phenotypes include clinical features such as neurologic deterioration, growth retardation, visceromegaly, and seizures in a severe early form; coarse facial features, angiokeratoma corporis diffusum, spasticity and delayed psychomotor development in a longer surviving form; and an unusual spondylometaphyseoepiphyseal dysplasia in yet another form
VRK2	0.00306	N/A	INTRON	his gene encodes a member of the vaccinia-related kinase (VRK) family of serine/threonine protein kinases. This gene is widely expressed in human tissues and has increased expression in actively dividing cells, such as those in testis, leukocytes, fetal liver, and carcinomas. Its protein localizes to the endoplasmic reticulum and has been shown to phosphorylate casein and undergo autophosphorylation. While several transcript variants may exist for this gene, the full-length nature of only one has been biologically validated to date
TRIM22	0.00326	N/A	INTRON	The protein encoded by this gene is a member of the tripartite motif (TRIM) family. The TRIM motif includes three zinc-binding domains, a RING, a B-box type 1 and a B-box type 2, and a coiled-coil region. This protein localizes to the cytoplasm and its expression is induced by interferon. The protein down-regulates transcription from the HIV-1 LTR promoter region, suggesting that function of this protein may be to mediate interferon's antiviral effects.
MITF	0.00347	N/A	INTRON	This gene encodes a transcription factor that contains both basic helix-loop-helix and leucine zipper structural features. It regulates the differentiation and development of melanocytes retinal pigment epithelium and is also responsible for pigment cell-specific transcription of the melanogenesis enzyme genes. Heterozygous mutations in the this gene cause auditory-pigmentary syndromes, such as Waardenburg syndrome type 2 and Tietz syndrome. Alternatively spliced transcript variants encoding different isoforms have been identified.

DHX36	0.00400	N/A	INTRON	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this DEAD box protein family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. The function of this gene product which is a member of this family, has not been determined.
RAD51	0.00427	N/A	INTRON	The protein encoded by this gene is a member of the RAD51 protein family. RAD51 family members are highly similar to bacterial RecA and Saccharomyces cerevisiae Rad51, and are known to be involved in the homologous recombination and repair of DNA. This protein can interact with the ssDNA-binding protein RPA and RAD52, and it is thought to play roles in homologous pairing and strand transfer of DNA. This protein is also found to interact with BRCA1 and BRCA2, which may be important for the cellular response to DNA damage. BRCA2 is shown to regulate both the intracellular localization and DNA-binding ability of this protein. Loss of these controls following BRCA2 inactivation may be a key event leading to genomic instability and tumorigenesis. Two alternatively spliced transcript variants of this gene, which encode distinct proteins, have been reported. Transcript variants utilizing alternative polyA signals exist.
TSPAN5	0.00433	N/A	INTRON	The protein encoded by this gene is a member of the transmembrane 4 superfamily, also known as the tetraspanin family. Most of these members are cell-surface proteins that are characterized by the presence of four hydrophobic domains. The proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth and motility.
DDX20	0.00514	N/A	INTRON	DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Based on their distribution patterns, some members of this family are believed to be involved in embryogenesis, spermatogenesis, and cellular growth and division. This gene encodes a DEAD box protein, which has an ATPase activity and is a component of the survival of motor neurons (SMN) complex. This protein interacts directly with SMN, the spinal muscular atrophy gene product, and may play a catalytic role in the function of the SMN complex on RNPs.

MUTED	0.00540	N/A	INTRON	This gene encodes a component of BLOC-1 (biogenesis of lysosome-related organelles complex 1). Components of this complex are involved in the biogenesis of organelles such as melanosomes and platelet-dense granules. A mouse model for Hermansky-Pudlak Syndrome is mutated in the murine version of this gene. Some transcripts of the downstream gene TXNDC5 overlap this gene, but they do not contain an open reading frame for this gene.
SYT11	0.00561	N/A	INTRON	May be involved in Ca(2+)-dependent exocytosis of secretory vesicles through Ca(2+) and phospholipid binding to the C2 domain or may serve as Ca(2+) sensors in the process of vesicular trafficking and exocytosis. Integral membrane protein. In substantia nigra, observed in neuronal cell bodies and neurites. Found in the core of the Lewy bodies in the brain of sporadic Parkinson disease patients.
KNTC1	0.00574	N/A	UPSTREAM INTRON	This gene encodes a protein that is one of many involved in mechanisms to ensure proper chromosome segregation during cell division. Experimental evidence indicated that the encoded protein functioned in a similar manner to that of the Drosophila rough deal protein. High expression in testis.
MAP7	0.00575	N/A	INTRON	The product of this gene is a microtubule-associated protein that is predominantly expressed in cells of epithelial origin. Microtubule-associated proteins are thought to be involved in microtubule dynamics, which is essential for cell polarization and differentiation. This protein has been shown to be able to stabilize microtubules, and may serve to modulate microtubule functions. Studies of the related mouse protein also suggested an essential role in microtubule function required for spermatogenesis.
GPM6B	0.00636	N/A	INTRON	May be involved in neural development.
Tenr	0.00696	N/A	UPSTREAM	testis nuclear RNA-binding protein

PUBLICATIONS

1. Polavarapu, N., McDonald, J.F., Differential gene expression between humans and chimps is associated with retrotransposon variation (Submitted to *Science*)

2. Piriyapongsa, J., Polavarapu, N., Borodovsky, M., McDonald, J.F., Exonization of the LTR transposable elements in human genome, *BMC Genomics* (accepted)

3. Menendez, L., Walker, D., Matyunina, L.V., Dickerson, E.B., Bowen, N.J., Polavarapu, N., Benigno, B.B, McDonald, J.F., Identification of candidate methylation-responsive genes in ovarian cancer, *Molecular Cancer,* Volume 6, Issue 10, 2007

4. Polavarapu, N., Bowen, N.J., McDonald, J.F., Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses, *Genome Biology***,** 2006, 7(6):R51

5. Polavarapu, N., Bowen, N.J., McDonald, J.F., Newly identified families of Human Endogenous Retroviruses (HERVs), *Journal of Virology*, 2006 80: 4640-4642

6. Polavarapu, N., Navathe, S.B., Ramnarayanan, R., Haque, A.U., Sahay, S., Lui, Y., Investigation into biomedical literature classification using support vector machines, Proc IEEE Comput Syst Bioinform Conf. 2005; 366-74.

7. Polavarapu, N., Bowen, N.J., McDonald, J.F., Consensus sequence of human endogenous retrovirus HERV1, *Repbase Reports*, Volume 5, Issue 6, 2005

8. Polavarapu, N., Bowen, N.J., McDonald, J.F., Consensus sequence of human endogenous retrovirus HERV4, *Repbase Reports*, Volume 5, Issue 6, 2005

REFERENCES

UCSC Genome Bioinformatics web site [http://genome.ucsc.edu/], March/2005.

Chimpanzee tRNA Database [http://lowelab.ucsc.edu/GtRNAdb/Ptrog/], January/2005.

Ensembl Chimp Genome [http://www.ensembl.org/Pan_troglodytes/index.html], September/2004.

Entrez Protein Database [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein], March/2005.

Chimpanzee Genome Browser [http://www.ensembl.org/Pan_troglodytes/], November/2004.

Ackerman H, Udalova I, Hull J, Kwiatkowski D (2002) Evolution of a polymorphic regulatory element in interferon-gamma through transposition and mutation. Mol Biol Evol 19(6): 884-890.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3): 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17): 3389-3402.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297(5585): 1301-1310.

Banville D, Boie Y (1989) Retroviral long terminal repeat is the promoter of the gene encoding the tumor-associated calcium-binding protein oncomodulin in the rat. J Mol Biol 207(3): 481-490.

Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK et al. (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol 9(16): 861-868.

Baumruker T, Gehe C, Horak I (1988) Insertion of a retrotransposon within the 3' end of a mouse gene provides a new functional polyadenylation signal. Nucleic Acids Res 16(15): 7241-7251.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J et al. (2004) Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A 101(14): 4894-4899.

Benit L, Dessen P, Heidmann T (2001) Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. J Virol 75(23): 11709-11719.

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes:m microcolinearity and its many exceptions. Plant Cell 12(7): 1021-1029.

Berquin IM, Ahram M, Sloane BF (1997) Exon 2 of human cathepsin B derives from an Alu element. FEBS Lett 419(1): 121-123.

Best S, Le Tissier P, Towers G, Stoye JP (1996) Positional cloning of the mouse retrovirus restriction gene Fv1. Nature 382(6594): 826-829.

Bird AP (1995) Gene number, noise reduction and biological complexity. Trends Genet 11(3): 94-100.

Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A 100(22): 13013-13018.

Boeke JD, Stoye JP (1997) Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. In: Coffin JM, Hughes SH, Varmus H, editors. Retroviruses. Plainview, NY: Cold Spring Harbor Laboratory Press. pp. 343-435.

Boissinot S, Furano AV (2005) The recent evolution of human L1 retrotransposons. Cytogenet Genome Res 110(1-4): 402-406.

Bolton EC, Boeke JD (2003) Transcriptional interactions between yeast tRNA genes, flanking genes and Ty elements: a genomic point of view. Genome Res 13(2): 254-263.

Bowen NJ, McDonald JF (1999) Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements. Genome Res 9(10): 924-935.

Bowen NJ, McDonald JF (2001) Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 11(9): 1527-1540.

Bowen NJ, Jordan IK (2002) Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol 4(3): 65-76.

Bowen NJ, Jordan IK, Epstein JA, Wood V, Levin HL (2003) Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of Schizosaccharomyces pombe. Genome Res 13(9): 1984-1997.

Britten RJ (1996) DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A 93(18): 9374-9377.

Britten RJ (1997) Mobile elements inserted in the distant past have taken on important functions. Gene 205(1-2): 177-182.

Britten RJ (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc Natl Acad Sci U S A 99(21): 13633-13635.

Brosius J (1999a) Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107(1-3): 209-238.

Brosius J (1999b) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238(1): 115-134.

Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A et al. (2002) A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418(6894): 145-151.

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L et al. (2003) Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci U S A 100(22): 13030-13035.

Carroll SB (2003) Genetics and the making of Homo sapiens. Nature 422(6934): 849-857.

Chaimanee Y, Jolly D, Benammi M, Tafforeau P, Duzer D et al. (2003) A Middle Miocene hominoid from Thailand and orangutan origins. Nature 422(6927): 61-65.

Chang-Yeh A, Mold DE, Huang RC (1991) Identification of a novel murine IAP-promoted placenta-expressed gene. Nucleic Acids Res 19(13): 3667-3672.

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68(2): 444-456.

Consortium CeS (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282(5396): 2012-2018.

Consortium CgS (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432(7018): 695-716.

Conte C, Dastugue B, Vaury C (2002) Coupling of enhancer and insulator properties identified in two retrotransposons modulates their mutagenic impact on nearby genes. Mol Cell Biol 22(6): 1767-1777.

Costas J, Naveira H (2000) Evolutionary history of the human endogenous retrovirus family ERV9. Mol Biol Evol 17(2): 320-330.

Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E et al. (2002) A single p450 allele associated with insecticide resistance in Drosophila. Science 297(5590): 2253-2256.

Dalmau J, Gultekin SH, Voltz R, Hoard R, DesChamps T et al. (1999) Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. Brain 122 (Pt 1): 27-39.

Daly TM, Rafii A, Martin RA, Zehnbauer BA (2000) Novel polymorphism in the FMR1 gene resulting in a "pseudodeletion" of FMR1 in a commonly used fragile X assay. J Mol Diagn 2(3): 128-131.

Dasilva C, Hadji H, Ozouf-Costaz C, Nicaud S, Jaillon O et al. (2002) Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the Tetraodon nigroviridis genome. Proc Natl Acad Sci U S A 99(21): 13636-13641.

DeBarry JD, Ganko EW, McCarthy EM, McDonald JF (2006) The contribution of LTR retrotransposon sequences to gene evolution in Mus musculus. Mol Biol Evol 23(3): 479-481.

Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67(3): 183-193.

Di Cristofano A, Strazullo M, Longo L, La Mantia G (1995) Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. Nucleic Acids Res 23(15): 2823-2830.

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757): 601-603.

Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C et al. (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A 102(3): 725-730.

Eberhard W (1985) Sexual selection in animal genitalia: Harvard University Press.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F et al. (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296(5566): 340-343.

Feuchter AE, Freeman JD, Mager DL (1992) Strategy for detecting cellular transcripts promoted by human endogenous long terminal repeats: identification of a novel gene (CDC4L) with homology to yeast CDC4. Genomics 13(4): 1237-1246.

Finnegan DJ (1992) Transposable elements. Curr Opin Genet Dev 2(6): 861-867.

Flavell RB (1986) Repetitive DNA and chromosome evolution in plants. Philos Trans R Soc Lond B Biol Sci 312(1154): 227-242.

Frazer KA, Chen X, Hinds DA, Pant PV, Patil N et al. (2003) Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. Genome Res 13(3): 341-346.

Friesen N, Brandes A, Heslop-Harrison JS (2001) Diversity, origin, and distribution of retrotransposons (gypsy and copia) in conifers. Mol Biol Evol 18(7): 1176-1188.

Gahan LJ, Gould F, Heckel DG (2001) Identification of a gene associated with Bt resistance in Heliothis virescens. Science 293(5531): 857-860.

Ganko EW, Fielman KT, McDonald JF (2001) Evolutionary history of Cer elements and their impact on the C. elegans genome. Genome Res 11(12): 2066-2074.

Garfinkel DJ (1997) Genetic loose change: how retroelements and reverse transcriptase heal broken chromosomes. Trends Microbiol 5(5): 173-175.

Gerber A, O'Connell MA, Keller W (1997) Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. Rna 3(5): 453-463.

Gerlo S, Davis JR, Mager DL, Kooijman R (2006) Prolactin in man: a tale of two promoters. Bioessays 28(10): 1051-1055.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428(6982): 493-521.

Goodchild NL, Wilkinson DA, Mager DL (1992) A human endogenous long terminal repeat provides a polyadenylation signal to a novel, alternatively spliced transcript in normal placenta. Gene 121(2): 287-294.

Goodman M (1999) The genomic record of Humankind's evolutionary roots. Am J Hum Genet 64(1): 31-39.

Goodwin TJ, Poulter RT (2000) Multiple LTR-retrotransposon families in the asexual yeast Candida albicans. Genome Res 10(2): 174-191.

Green MM (1988) Mobile DNA elements and spontaneous gene mutation. In Eukaryotic traosposable elements as mutagenic agents. Cold Spring Harbor laboratory, Cold Spring Harbor, NY: 41-50.

Griffiths DJ (2001) Endogenous retroviruses in the human genome sequence. Genome Biol 2(6): REVIEWS1017.

Gu J, Gu X (2003) Induced gene expression in human brain after the split from chimpanzee. Trends Genet 19(2): 63-65.

Hamdi HK, Nishio H, Tavis J, Zielinski R, Dugaiczyk A (2000) Alu-mediated phylogenetic novelties in gene regulation and development. J Mol Biol 299(4): 931-939.

Han K, Konkel MK, Xing J, Wang H, Lee J et al. (2007) Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science 316(5822): 238-240.

Harendza CJ, Johnson LF (1990) Polyadenylylation signal of the mouse thymidylate synthase gene was created by insertion of an L1 repetitive element downstream of the open reading frame. Proc Natl Acad Sci U S A 87(7): 2531-2535.

Hasler J, Strub K (2006) Alu elements as regulators of gene expression. Nucleic Acids Res 34(19): 5491-5497.

Hickey DA (1982) Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101(3-4): 519-531.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298(5591): 129-149.

Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A et al. (2002) Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol 3(12): RESEARCH0085.

Hu MC, Qiu WR, Wang X, Meyer CF, Tan TH (1996) Human HPK1, a novel human hematopoietic progenitor kinase that activates the JNK/SAPK kinase cascade. Genes Dev 10(18): 2251-2264.

Hughes JF, Coffin JM (2005) Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. Genetics 171(3): 1183-1194.

Illarionova AE, Vinogradova TV, Sverdlov ED (2007) Only those genes of the KIAA1245 gene subfamily that contain HERV(K) LTRs in their introns are transcriptionally active. Virology 358(1): 39-47.

Iwashita S, Osada N, Itoh T, Sezaki M, Oshima K et al. (2003) A transposable element-mediated gene divergence that directly produces a novel type bovine Bcnt protein including the endonuclease domain of RTE-1. Mol Biol Evol 20(9): 1556-1563.

Jiang N, Bao Z, Temnykh S, Cheng Z, Jiang J et al. (2002) Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. Genetics 161(3): 1293-1305.

Johnson WE, Coffin JM (1999) Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci U S A 96(18): 10254-10260.

Jordan IK, McDonald JF (2002) A Biologically Active Family of Human Endogenous Retroviruses Evolved from an Ancient Inactive Lineage. Genome Letters 1(3): 1-5.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19(2): 68-72.

Jukes TH, Cantor CR, editors (1969) Evolution of protein molecules. New York: Academic Press. 21 - 132 p.

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16(9): 418-420.

Kaiser SM, Malik HS, Emerman M (2007) Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. Science 316(5832): 1756-1758.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R et al. (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol 3(12): RESEARCH0084.

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42(1): 59-65.

Kapitonov VV, Jurka J (1999) The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J Mol Evol 48(2): 248-251.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32(Database issue): D493-496.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31(1): 51-54.

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33(1): 102-106.

Kazazian HH, Jr. (1998) Mobile elements and disease. Curr Opin Genet Dev 8(3): 343-350.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M et al. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309(5742): 1850-1854.

Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution 55(1): 1-24.

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res 8(5): 464-478.

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188(4184): 107-116.

Kreahling J, Graveley BR (2004) The origins and implications of Aluternative splicing. Trends Genet 20(1): 1-4.

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33: 479-532.

Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5(2): 150-163.

Kunimatsu Y, Ratanasthien B, Nakaya H, Saegusa H, Nagaoka S (2004) Earliest Miocene hominoid from Southeast Asia. Am J Phys Anthropol 124(2): 99-108.

Laimins L, Holmgren-Konig M, Khoury G (1986) Transcriptional "silencer" element in rat repetitive sequences associated with the rat insulin 1 gene locus. Proc Natl Acad Sci U S A 83(10): 3151-3155.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822): 860-921.

Landry JR, Rouhi A, Medstrand P, Mager DL (2002) The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol 19(11): 1934-1942.

Larkin DM, Everts-van der Wind A, Rebeiz M, Schweitzer PA, Bachman S et al. (2003) A cattle-human comparative map built with cattle BAC-ends and human genome sequence. Genome Res 13(8): 1966-1972.

Lavie L, Medstrand P, Schempp W, Meese E, Mayer J (2004) Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. J Virol 78(16): 8788-8798.

Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in Arabidopsis thaliana. Proc Natl Acad Sci U S A 97(13): 7376-7381.

Leeton PR, Smyth DR (1993) An abundant LINE-like element amplified in the genome of Lilium speciosum. Mol Gen Genet 237(1-2): 97-104.

Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409(6822): 847-849.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069): 803-819.

Ling J, Pi W, Bollag R, Zeng S, Keskintepe M et al. (2002) The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. J Virol 76(5): 2410-2423.

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V et al. (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. Science 276(5312): 561-567.

Liu AY, Abraham BA (1991) Subtractive cloning of a hybrid human endogenous retrovirus and calbindin gene in the prostate cell line PC3. Cancer Res 51(15): 4107-4110.

Liu G, Zhao S, Bailey JA, Sahinalp SC, Alkan C et al. (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res 13(3): 358-368.

Llorens C, Marin I (2001) A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 18(8): 1597-1600.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5): 955-964.

Mager DL (1989) Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H. Virology 173(2): 591-599.

Mager DL, Hunter DG, Schertzer M, Freeman JD (1999) Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics 59(3): 255-263.

Makalowski W (2000) Genomic scrap yard: how genomes utilize all that junk. Gene 259(1-2): 61-67.

Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN et al. (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2(1): e2.

Malik HS, Henikoff S (2005) Positive selection of Iris, a retroviral envelope-derived host gene in Drosophila melanogaster. PLoS Genet 1(4): e44.

Matsumine H, Herbst MA, Ou SH, Wilson JD, McPhaul MJ (1991) Aromatase mRNA in the extragonadal tissues of chickens with the henny-feathering trait is derived from a distinctive promoter structure that contains a segment of a retroviral long terminal repeat. Functional organization of the Sebright, Leghorn, and Campine aromatase genes. J Biol Chem 266(30): 19900-19907.

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19(3): 362-367.

McCarthy EM, McDonald JF (2004) Long terminal repeat retrotransposons of Mus musculus. Genome Biol 5(3): R14.

McCarthy EM, Liu J, Lizhi G, McDonald JF (2002) Long terminal repeat retrotransposons of Oryza sativa. Genome Biol 3(10): RESEARCH0053.

McClintock B (1946) Maize genetics. (45): 176-186.

McClintock B (1948) Mutable loci in maize. Carnegie Institute of Washington Year Book 47: 155-169.

McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16: 13-47.

McClintock B (1984) The significance of responses of the genome to challenge. Science 226(4676): 792-801.

McCollum AM, Ganko EW, Barrass PA, Rodriguez JM, McDonald JF (2002) Evidence for the adaptive significance of an LTR retrotransposon sequence in a Drosophila heterochromatic gene. BMC Evol Biol 2: 5.

McDonald JF (1993) Evolution and consequences of transposable elements. Curr Opin Genet Dev 3(6): 855-864.

McDonald JF (1995) Transposable elements: possible catalysts of organismic evolution. Trends Ecol Evol 10: 123-126.

McDonald JF (1998) Transposable elements, gene silencing and macroevolution. Trends Ecol Evol 13: 94-95.

Medstrand P, Mager DL (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. J Virol 72(12): 9782-9787.

Medstrand P, Landry JR, Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J Biol Chem 276(3): 1896-1903.

Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D et al. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res 110(1-4): 342-352.

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11(10): 1660-1676.

Mi S, Lee X, Li X, Veldman GM, Finnerty H et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403(6771): 785-789.

Mikkelsen T, Hillier LW, Eichler EE, Zody MC, David JB et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055): 69-87.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447(7141): 167-177.

Miller WJ, McDonald JF, Pinsker W (1997) Molecular domestication of mobile elements. Genetica 100(1-3): 261-270.

Miller WJ, McDonald JF, Nouaud D, Anxolabehere D (1999) Molecular domestication--more than a sporadic episode in evolution. Genetica 107(1-3): 197-207.

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD et al. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31(2): 159-165.

Murnane JP, Morales JF (1995) Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. Nucleic Acids Res 23(15): 2837-2839.

Nagasaki K, Schem C, von Kaisenberg C, Biallek M, Rosel F et al. (2003) Leucine-zipper protein, LDOC1, inhibits NF-kappaB activation and sensitizes pancreatic cancer cells to apoptosis. Int J Cancer 105(4): 454-458.

Nakamura TM, Morin GB, Chapman KB, Weinrich SL, Andrews WH et al. (1997) Telomerase catalytic subunit homologs from fission yeast and human. Science 277(5328): 955-959.

Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17(11): 619-621.

Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M et al. (2005) A genome-wide survey of structural variation between human and chimpanzee. Genome Res 15(10): 1344-1356.

Nigumann P, Redik K, Matlik K, Speek M (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79(5): 628-634.

Oda H, Nakabeppu Y, Furuichi M, Sekiguchi M (1997) Regulation of expression of the human MTH1 gene encoding 8-oxo-dGTPase. Alternative splicing of transcription products. J Biol Chem 272(28): 17843-17850.

Ono R, Nakamura K, Inoue K, Naruse M, Usami T et al. (2006) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38(1): 101-106.

Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284(5757): 604-607.

Pardue ML, Danilevskaya ON, Lowenhaupt K, Slot F, Traverse KL (1996) Drosophila telomeres: new views on chromosome evolution. Trends Genet 12(2): 48-52.

Paulson KE, Matera AG, Deka N, Schmid CW (1987) Transcription of a human transposon-like sequence is usually directed by other promoters. Nucleic Acids Res 15(13): 5199-5215.

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8): 2444-2448.

Polavarapu N, Bowen NJ, McDonald JF (2006a) Newly identified families of Human Endogenous Retroviruses (HERVs). J Virol 80(9).

Polavarapu N, Bowen NJ, McDonald JF (2006b) Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol 7(6): R51.

Richter TE, Ronald PC (2000) The evolution of disease resistance genes. Plant Mol Biol 42(1): 195-204.

Rostoks N, Park YJ, Ramakrishna W, Ma J, Druka A et al. (2002) Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. Funct Integr Genomics 2(1-2): 51-59.

Rothenburg S, Eiben M, Koch-Nolte F, Haag F (2002) Independent integration of rodent identifier (ID) elements into orthologous sites of some RT6 alleles of Rattus norvegicus and Rattus rattus. J Mol Evol 55(3): 251-259.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274(5288): 765-768.

Satoh N, Satou Y, Davidson B, Levine M (2003) Ciona intestinalis: an emerging model for whole-genome analyses. Trends Genet 19(7): 376-381.

Schwartz JH (1984) The evolutionary relationships of man and orang-utans. Nature 308(5959): 501-505.

Selker EU, Tountas NA, Cross SH, Margolin BS, Murphy JG et al. (2003) The methylated component of the Neurospora crassa genome. Nature 422(6934): 893-897.

Shapiro J (1977) DNA insertion elements and the evolution of chromosome primary structure. Trends Biochem Sci 2: 622-627.

Smit AF (1993) Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res 21(8): 1863-1872.

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9(6): 657-663.

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12(7): 1060-1067.

Speek M (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21(6): 1973-1985.

Stavenhagen JB, Robins DM (1988) An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene. Cell 55(2): 247-254.

Stoye JP (2001) Endogenous retroviruses: still active after all these years? Curr Biol 11(22): R914-916.

Sverdlov ED (2000) Retroviruses and primate evolution. Bioessays 22(2): 161-171.

Tan KO, Tan KM, Chan SL, Yee KS, Bevort M et al. (2001) MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. J Biol Chem 276(4): 2802-2807.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25(24): 4876-4882.

Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365: 104-110.

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev 6(8): 1457-1465.

Tomasello M, Call J (1997): Oxford Univ. Press, Oxford.

Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 74(8): 3715-3730.

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK et al. (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11(19): 1531-1535.

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19(10): 530-536.

van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL (2005) Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res 15(9): 1243-1249.

Varki A, Altheide TK (2005) Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res 15(12): 1746-1758.

Vicient CM, Jaaskelainen MJ, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125(3): 1283-1292.

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28(9): 913-922.

Volff JN, Korting C, Froschauer A, Sweeney K, Schartl M (2001) Non-LTR retrotransposons encoding a restriction enzyme-like endonuclease in vertebrates. J Mol Evol 52(4): 351-360.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915): 520-562.

Wills NM, Moore B, Hammer A, Gesteland RF, Atkins JF (2006) A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. J Biol Chem 281(11): 7082-7088.

WoldeGabriel G, Haile-Selassie Y, Renne PR, Hart WK, Ambrose SH et al. (2001) Geology and palaeontology of the Late Miocene Middle Awash valley, Afar rift, Ethiopia. Nature 412(6843): 175-178.

Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403(6767): 304-309.

Xiong Y, Eickbush TH (1988) Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol 5(6): 675-690.

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. Embo J 9(10): 3353-3362.

Yamada T, Ohtani S, Sakurai T, Tsuji T, Kunieda T et al. (2006) Reduced expression of the endothelin receptor type B gene in piebald mice caused by insertion of a retroposon-like element in intron 1. J Biol Chem 281(16): 10799-10807.

Yang Z, Boffelli D, Boonmark N, Schwartz K, Lawn R (1998) Apolipoprotein(a) gene enhancer resides within a LINE element. J Biol Chem 273(2): 891-897.

Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P et al. (2005) Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. PLoS Biol 3(4): e110.

Yu F, Zingler N, Schumann G, Stratling WH (2001) Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. Nucleic Acids Res 29(21): 4493-4501.

Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res 33(3): 946-954.

Zhou YH, Zheng JB, Gu X, Saunders GF, Yung WK (2002) Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. Genome Res 12(11): 1716-1722.